In Pursuit of Regulatable LLMs

Eoin M. Kenny
Massachusetts Institute of Technology
Cambridge, MA, U.S.A.
ekenny@mit.edu

Julie A. Shah
Massachusetts Institute of Technology
Cambridge, MA, U.S.A.
julie_a_shah@csail.mit.edu

Abstract

Large-Language-Models (LLMs) are arguable the biggest breakthrough in artificial intelligence to date. Recently, they have come to the public Zeitgeist with a surge of media attention surrounding ChatGPT, a large generative language model released by OpenAI which quickly became the fastest growing application in history. This model achieved unparalleled human-AI conversational skills, and even passed various mutations of the famous Turing test which measures if AI systems have achieved general intelligence. Naturally, the world at large wants to utilize these systems for various applications, but in order to do-so in truly sensitive domains, the models must often be regulatable in order to be legally used. In this short paper, we propose one approach towards such systems by forcing them to reason using a combination of (1) human-defined concepts, (2) Case-Base Reasoning (CBR), and (3) counterfactual explanations. All of these have support in user testing that they are understandable and useful to practitioners of AI systems, moreover counterfactuals have been argued as compliant with the GDPR. We envision this approach will be able to provide more transparent LLMs for text classification tasks and be fully regulatable and auditable.

1 Introduction

Perhaps the most important breakthrough in recent ML research is that of Large-Language-Models (LLMs) [34], these systems have seemingly mastered the nuances of human conversation and have even been shown to closely follow some of the semantic rules innate to human language understanding [32]. However, these systems cannot escape the same core issue that underlines most neural network architectures, in that they are black-boxes with no obvious interpretable decision-making process, making it impossible to trust them in practice for any sensitive application [29]. To combat this, Explainable AI (XAI) has become a vast research field [28, 16, 9, 13], with massive potential to e.g. make models auditable, debug self-driving cars, and calibrate appropriate trust between humans and AI in high stakes scenarios [31]. However, to date there is no convincing example of XAI being used to make regulatable LLMs, a deficit we address in this short paper. In doing so, we hope to stimulate an interesting conversation surrounding this topic.

Our core motivation lies in the fact that many institutions require their employees (and by extension their models) to use specific concepts in sensitive decisions, but due to the black-box nature of LLMs, there is absolutely no way to verify this is actually happening [29, 18]. Hence, we posit that in order to be auditable and regulatable, *we must force these models to use specific human-defined concepts in their inference process*. Moreover, they need to be used in an understandable way, which further motivates our usage of Case-Based Reasoning (CBR) to visualize the usage of these concepts, which has shown to be understandable by end-users [17, 21, 36], useful to help decision making [7], and preferred as as form of explanation over other popular approaches [10]. To the best of our knowledge,

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

this idea to force the usage of human-defined concepts in CBR is a novel research direction, and moreover it is seen as one of the grand challenges of XAI [30].

2 Related Work

XAI has been prominent in every area of Machine Learning (ML) [9, 16, 13, 25], but to date, there has been somewhat of a void in natural language processing. To stay relevant, we limit our discussion here to work which uses concepts, CBR, or causality (for counterfactual explanations). Moreover, we focus on classification-based XAI methods which are "interpretable by design" (as opposed to post-hoc explanations [28, 2]), since we believe this will likely be a necessary prerequisite for making them regulatable [29, 37].

Concept-based XAI methods for LLMs strive to use "human understandable concepts" in their inference process [22]. In LLMs, the idea is to typically learn "black-box" concepts automatically via text rationales [5], and have humans label what they are once the system in trained. Once this is done, the concepts are typically fed into a linear model to produce an interpretable prediction. Notable work in this area was completed by Bouchacourt & Denoyer [3], but it suffers from poor performance. Antognini & Faltings [1] proposed CONRAT, a similar method which used *multiple* concepts in its decisions, and is reported to perform better in terms of accuracy. In contrast to these approaches, we propose to (1) get humans to define the important concepts *a-priori*, and (2) visualize these concept-based explanations with CBR, so that they are more understandable and guaranteed to be legally compliant.

CBR for interpretable LLMs is a fairly new idea, it strives to use real examples from the training data directly in inference, so that explanations may be parsed as e.g. *"I think this test instance is class c, because it is similar to this training instance which was also class c"*. Notable work in this area can be traced back to Ming et al. [26], but their approach only works for recurrent neural networks. Das et al. [8] proposed ProtoTEx, which classifies test instances with reference to learned prototypes (i.e., examples). Notably however, this work uses whole training examples for similarity, rather than concepts or sentences which would give more detail. Van Aken et al. [33] built upon this by proposing ProtoPPatient, which works for multi-label classification. However, once again, all of these methods cannot get humans to define the important concepts to be used in the CBR process.

With regards to causality [27], it is argued that counterfactual explanations are GDPR compliant [35], and causality is likely necessary to guarantee good counterfactuals [27, 13]. Counterfactuals show an imagined alternative to the current situation in which the user's classification changes with e.g. *"if you double your medication, your chest pain will likely dissapear"*. In tabular data, causality has largely been implemented with structural causal models (SCMs) [12, 15, 14], but to the best of our knowledge, using SCMs for explanations has not been attempted in LLMs (or even deep learning in general). In this work, we aim to accomplish this by training the LLM to identify the presence of human-defined concepts, and then use these as features in the inference process. By doing this, we can use the SCM to generate a counterfactual explanation.

Perhaps the most closely related work to what we propose here is that of Kenny et al. [20]. Specifically, the authors propose 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, most likely because they only use a single example to represent each concept. We build upon this work by (1) collecting a large human-annotated dataset for each concept to avoid over-fitting, (2) adding an SCM to allow counterfactual explanations, and (3) adapting the framework for LLMs. With regards to point (3), this is important as Kenny et al. [20] only worked with "whole" cases for explanations, whilst in our case with LLMs we are breaking each case down into individual sentence embeddings (see Figure 1), which is a far more intricate learning process.

3 Proposed Method

In the model shown in Figure 1, a test instance, x, is mapped to a latent representation, z, via the original encoder network: $z = f_{\text{enc}}(x)$. This vector z can then be further subdivided into a separate embedding z_i for each sentence in the input text, giving a set of embeddings $Z \in \{z_i\}_{i=1}^{n}$. This set is then passed into e.g. the first transformation h_1 to measure each sentence's similarity to all the

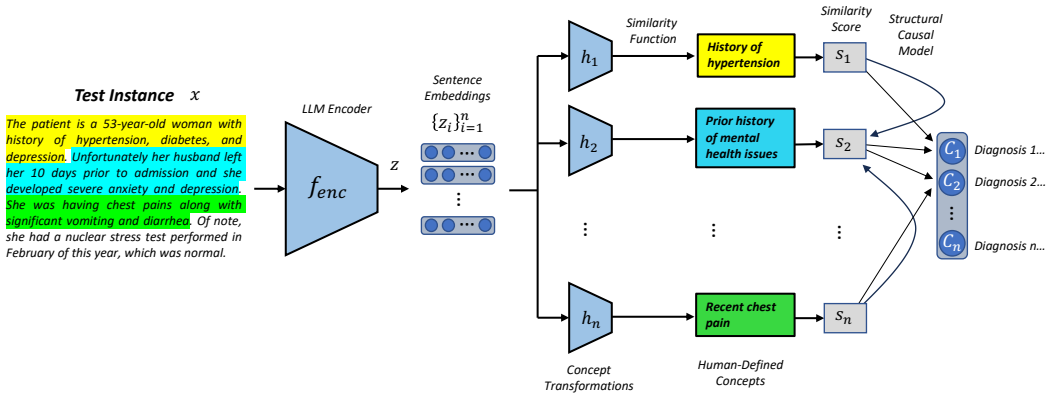

Figure 1: Proposed Framework: First, a test instance x is encoded by the LLM into sentence embeddings z_i. Then, each z_i is passed through every transformation $h_{i,j}$ and has its similarity measured to the labelled examples of each concept. The sentence which activates most for each concept outputs a similarity score, which will be high if they are very similar (such as the first sentence and the concept of "History of Hypertension"), and low if they are not similar at all (e.g., the last sentence didn't activate any concept notably). Note again each concept is a collection of labelled examples for each concept. The similarity feature vector is then classified with e.g. a linear model (not pictured above). If desired, the SCM can also be used to generate a counterfactual explanation.

labelled examples of the concept. The maximum similarity score is then used as the feature value for the concept. Moreover, this labelled example used to calculate the maximum similarity score can also be used as the explanation for classifying that particular concept by saying e.g. *"I think this sentence has the concept 'History of Hypertension' because it is very similar to this labelled example [...show example to user...] stored for this concept"*.

3.1 Training Procedure

Each concept is represented by a set of "prototypical cases", which in practice are human defined examples of the concept that must be annotated (or perhaps generated/clustered if labelled examples are difficult to acquire); in general, the more there are, the better the results should be. So, after each transformation h, there are a set of labelled examples of the concept, and the distance of the embedding z_i to the closest example is used in the similarity function to output a similarity score for that concept (but in the figure we simply write the concept description – e.g. "History of Hypertension" for simplicity and clarity). The training process may be summarized as:

1. Identify the human-friendly concepts in the domain desired for the model to use in its decisions.
2. Acquire a labelled dataset of these concepts (note this may be AI generated or clustered).
3. **Forward Pass:** Pass the sentence embeddings of x into the first h_i, alongside human annotated examples, and find the sentence in x which is most similar to one of the annotated examples. Repeat this for all h_i to get a similarity output for all concepts, and run the feature values through an output model (e.g., a linear one) to get a prediction.
4. Train this either (1) end-to-end, or (2) as a fine-tuning process using a pre-trained LLM encoder (either frozen or not).

Notably, this will require passing a large amount of labelled data through the network at each iteration (i.e., the normal training data and concept data), which may causes memory issues, but this can be circumvented by limiting the amount of labelled concept data each forward pass to a subset of the data (which should also act as a form of regularization).

Another more pressing issue, is that if we constrain the model to use specific concepts, it may compromise performance. There are several things to consider here: Firstly, just because a model

is less accurate, it doesn't mean it's actually worse, the better model could be relying on spurious correlations. Secondly, if the defined concepts are high-level, they may be general enough to not compromise performance, likewise if they are very specific, it may overly constrain the model. Thirdly, alongside the human-defined concepts, we could also learn a black-box concept, which is trained as a residual to maintain performance [38], although this would require manually labelling of this concept post-training by observing instances in this area of the function.

3.2 Explanations

Due to the "interpretable-by-design" nature of the model, it is very straightforward to offer both causal and counterfactual explanations.

Causal Explanation These explanations are naturally generated by the models inference process due to it using CBR. For example, to explain the classification of e.g. the concept "History of Hypertension" in the first sentence, you may say *"I think the first sentence shows the patient has a history of hypertension, because it is similar to a sentence in a previous patient [...show example...] who also had this"*. This explanation process can be repeated to explain the classification of all concepts. This CBR-type explanation shows *why* the model identified the presence of the concept [24, 16, 6], in that it is essentially saying *"This sentence is similar to this labelled example, hence I think they are the same concept"*. Moreover, because the concepts are human-defined, they are interpretable and should (in many domains) make the model more regulatable. Finally, as these are then passed through an output linear model, the final prediction is interpretable.

Counterfactual Explanation Importantly, due to the SCM, a counterfactual may also be generated which could e.g. modify the feature chest pain, whilst automatically (thanks to the SCM) mutating causally dependent features; for example it may say *"If you got rid of your chest pain, it would also remove part of your depression"*. Note that the concept of "recent chest pain" is causally linked to "prior history of mental health issues", so altering the former affects the latter, which is the point of having the SCM (i.e., to generate *only* plausible counterfactuals). Notably, this wouldn't require actually generating new text, which would help with plausibility [23, 19].

3.3 Preliminary Results

The proposed framework has been tested using entire test instances (i.e, "cases") as labelled concepts in a sensitive domain for which interpretability is critical.[1] Results showed that the interpretable model had the same accuracy as black-box counterparts, and delivered CBR explanations which generally aligned with domain experts about what constituted "similar looking" in the domain. Current tests are examining the potential of the framework to use individual sentences as concepts rather than the whole instance, early signs indicate this is promising. Lastly, we are working with domain experts to build SCMs by hand which can generate counterfactual explanations.

4 Discussion

For the past ten years at large, the public has grown particularly weary of big tech companies and governments increasingly monitoring their personal lives with (often) unsolicited data harvesting. The most notable example to combat this and win back the public trust was the famous General Data Protection Regulation (GDPR) which attempted to regulate ML in Europe. However, this far from solved the problem, and there are a host of new regulation laws coming in 2023 [39], with even the head of the company behind ChatGPT recently stating that ML should be tighter regulated [11]. As explanation is usually seen as a vital part of regulation [4], designing interpretable LLMs makes sense as a step forward here. In this short paper, we have proposed (and begun to test) a basic framework for LLMs which should allow them to be largely regulatable. We welcome feedback, comments, and criticisms from the community, and hope to stimulate an interesting discussion surrounding this pressing topic.

[1] Note that due to funding and legal constraints the authors are not permitted to discuss this further, but the domain in question is irrelevant to the purpose of this paper.

References

[1] D. Antognini and B. Faltings. Rationalization through concepts. *arXiv preprint arXiv:2105.04837*, 2021.

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015.

[3] D. Bouchacourt and L. Denoyer. Educe: Explaining model decisions through unsupervised concepts extraction. *arXiv preprint arXiv:1905.11852*, 2019.

[4] M. C. Buiten. Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation*, 10(1):41–59, 2019.

[5] A. Chan, S. Nie, L. Tan, X. Peng, H. Firooz, M. Sanjabi, and X. Ren. Frame: Evaluating simulatability metrics for free-text rationales. *arXiv preprint arXiv:2207.00779*, 2022.

[6] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, pages 8928–8939, 2019.

[7] V. Chen, Q. V. Liao, J. W. Vaughan, and G. Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *arXiv preprint arXiv:2301.07255*, 2023.

[8] A. Das, C. Gupta, V. Kovatchev, M. Lease, and J. J. Li. Prototex: Explaining model decisions with prototype tensors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

[9] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[10] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33:4211–4222, 2020.

[11] C. Kang. How sam altman stormed washington to set the a.i. agenda, Jun 2023.

[12] A.-H. Karimi, G. Barthe, B. Belle, and I. Valera. Model-agnostic counterfactual explanations for consequential decisions. *arXiv preprint arXiv:1905.11190*, 2019.

[13] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. 2021.

[14] A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

[15] A.-H. Karimi, J. Von Kügelgen, B. Schölkopf, and I. Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33:265–277, 2020.

[16] M. T. Keane and E. M. Kenny. How case-based reasoning explains neural networks: A theoretical analysis of xai using post-hoc explanation-by-example from a survey of ann-cbr twin-systems. In *International Conference on Case-Based Reasoning*, pages 155–171. Springer, 2019.

[17] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies. *Artificial Intelligence*, page 103459, 2021.

[18] E. M. Kenny and M. T. Keane. Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in xai. *Knowledge-Based Systems*, 233:107530, 2021.

[19] E. M. Kenny and M. T. Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11575–11585, 2021.

[20] E. M. Kenny, M. Tucker, and J. Shah. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *The Eleventh International Conference on Learning Representations*, 2023.

[21] B. Kim, C. Rudin, and J. A. Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, pages 1952–1960, 2014.

[22] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[23] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint arXiv:1907.09294*, 2019.

[24] D. B. Leake. Case-based reasoning: experiences, lessons, and future directions. 1996.

[25] Z. C. Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[26] Y. Ming, P. Xu, H. Qu, and L. Ren. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 903–913, 2019.

[27] J. Pearl. Causality: Models, reasoning and inference cambridge university press. *Cambridge, MA, USA,*, 9:10–11, 2000.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

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

[30] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

[31] L. Sanneman and J. A. Shah. 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, 2022.

[32] M. Tucker, P. Qian, and R. Levy. What if this modified that? syntactic interventions with counterfactual embeddings. Association for Computational Linguistics, 2021.

[33] B. Van Aken, J.-M. Papaioannou, M. Naik, G. Eleftheriadis, W. Nejdl, F. Gers, and A. Loeser. 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, 2022.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[36] G. Warren, B. Smyth, and M. T. Keane. "better" counterfactuals, ones people can understand: Psychologically-plausible case-based counterfactuals using categorical features for explainable ai (xai). In *International conference on case-based reasoning*, pages 63–78. Springer, 2022.

[37] T. Wischmeyer and T. Rademacher. *Regulating artificial intelligence*, volume 1. Springer, 2020.

[38] M. Yuksekgonul, M. Wang, and J. Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

[39] K. Zhu. The state of state ai laws: 2023.