Explainability for AI-based business digital business transformation (preprint)

Łukasz Górski[1] Shashishekar Ramakrishna[2]

[1]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Poland
[2]Research Advocacy and Management School of IPR, Bangalore, India

Abstract

Introduction of artificial intelligence-based methods, including Large Language Models is transforming the business as we see it. The legislators have, however, noted the introduction of black-box models puts limits to our ability to understand the decision made with the use of those sophisticated tools. In this paper, we will sketch how the newly drafted EU AI Act fits into this landscape and we will present our proof-of-concept system aimed at bringing explainability into the black box systems, including the ones based on LLMs.

1 Introduction

Recent introduction of new data- and computationally intensive artificial intelligence-based methods is accelearating current digital transformation of businesses. Advancements of in the area of natural language processing (NLP) became a factor that caused reevalution of strategies across companies. Whilst current emerging Large Language Models are not yet at the level of Artificial General Intelligence (AGI), they allow to develop natural language-based software that allows insight that bear striking similarity to human capabilities. On the other hand, the advancements of the Big Data processing and the high complexity of AI tools is necessarily connected with the increase of the opacity of such systems. One of the most severe societal consequences is the danger it puts to the right of end users and their ability to comprehend the prediction made by the AI-based system. This has also been acknowledged by public authorities on the European Union level, with the proposal of AI act aimed to protect the rights of citizens and increase transparency.

In this paper, we present a proof-of-concept system we have developed for the explainable Large Language Modelling. It is shown how such a system can

be employed for regulatory and legal compliance, and how it fits within the existing business procedures.

2 Regulatory and legal compliance

The adherence to legal standards (legal compliance) as well as to internal-corporate/industrial guidelines (regulatory compliance) is a necessary condition for the implementation of business processes. Compliance aims to join two distinct domains, the business practices, and the normative regulatory domain [7]. This is a multifacet endaevor, pertaining not only to the tasks to perform, but also their effects as well as artefacts generated while executing the tasks (e.g. the dataset developed) [4]. The explosive development of AI-based tools has made the European Union (EU) to double down on its citizen-protection measures, already exhibited by General Data Protection Regulation (GDPR). The introduction of European-level AI Act is meant to safeguard citizens interest in the wake of AI penetrating business practices. As in the case of GDPR, the regulators expect the AI act to impact businesses worldwide and do not construe its impact only to EU's territory. In this paper, we take the similar approach, noting that at least some of the provisions of the Act are general enough to be considered in an universal context. However, the European regulation is not the only one underway. China has already established a regulatory framework, USA and UK are also following suit [6].

In general, two strategies to compliance assessment are viable. In design-time compliance approach, the checks are introduced early in the process design stage, and the requirements are presented as a set of logic clauses that the designed process has to adherere to. In run-time compliance approach, the compliance reports are generated during the execution and those are produced by the specialized software [7]. Both of those approaches can be aligned with the framework offered by the EU AI Act.

Explainable Artificial Intelligence is a body of research curriculum and industrial practices that aims to make the results and inner-working of contemporary AI-based models accessible to human. The number of parameters that compose the today's system make it impossible to trace back the reasoning that caused a system to make a prediction. XAI aims to undercover this reasoning, by - for example - showing which parts of the input contributed to the output, what features are most important for a given prediction or how can an input sample be perturbed to invert the system's prediction.

In general, when talking about AI and how to make their results understandable, three terms are to be crucial. We talk about *interpretability*, when we discuss the inherent properties of a system, that allow it be understood without the reference to any external tools or methods. For example, if we had a system that's prediction, y, is lineary dependant on input variable x ($y = ax + b$), such system would be trivial to explain, simply by referring to the values of a and b coefficients. We would be talking about *explainable* systems, when we apply external methods to gain insights into their inner working and their rationale for

a given decision. For example, the SHAP method we refer to in the later part of this paper, can be recalled. Thirdly, in case of *justification*, we communicate why a given decision is good, we do not necessarily refer to mechanistic understanding of how it was arrived at [5]. Thus, the function here is persuasive. In legal contexts, such justification would not only refer the AI-based model, but its relevance to normative background.

Whilst whether the GDPR introduced a separate right to explanation is still under contention due to the vagueness of relevant provisions [8], the AI Act is clearer in this respect. It introduced a tiered approach, wherein different requirements are set forth, depending on the application and risks associated with it. Certain types of applications are forbidden (e.g. facial recognition for social credit scoring), the lowest risk is associated only with certain informative requirements for transparency (art. 52).

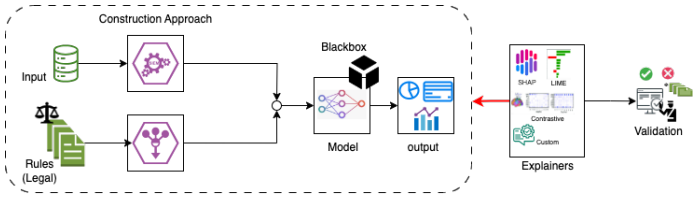
Most nuanced regulation is aimed at high risk applications, such as those pertaining to credit scoring, justice, or medical applications, which will be our focal point. Therein the act talks about the 'interpretability' that has to be provided, as in art. 13 and art. 14. Borrowing the conceptual framework from [5], we will view art. 13 as relating to *ex ante* explainability, and art. 14 - as pertaining to *ex post* explainability. Art. 13 can be connected with the prior creation of clean and understandable instructions for system's use. Art. 14 can be read as requiring *ex post* explainability from the system. After the prediction is made, a natural person should be able to oversee the system, whilst also using relevant intepretation tools and methods. The latter is the focal point of this paper.

Whilst the act itself was conceived as future-proof and general enough to tackle the technological change, its drafting saw the rise of GenAI, systems able to generate natural language, sounds, or pictures with human-like properties. Those amendments are, however, still under the discusion within the EU's institutions. For example, present in the European Parliment's proposal is art. 28b, that mandates, among other provisions, that foundation models (a broader category encompassing GenAI) should offer intepretability, and such interpretability should be assured by referencing, *inter alia*, model evaluation and testing. The methodology presented by us in this paper, can be used to assure this aspect.

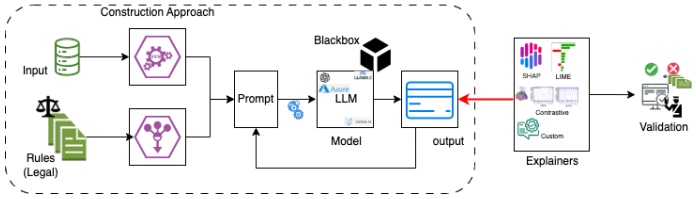
3 Towards explainable large language models

A high-level architecture depicting the meta-models for both classical AI systems backed by deep learning (DL) and generative AI (Gen.AI) is shown in 2. Both approaches start by guiding both the input data and the domain knowledge (in the form of rules) into a black-box system. In classical systems, the control of such black-boxes was achieved using weights and biases. However, with GenAI, we can tune the system behaviour with steering parameters and prompt design methodologies. The output from such systems can be requested in a format that is reasonable for the next system in the pipeline to consume.

An XAI system usually sits outside the prediction system. While some efforts



(a) Classical approach.



(b) Gen. AI approach.

Figure 1: High-level architecture showing the meta-models for an XAI system.

have been done to integrate the XAI methodologies into the main system, the main drawback is on the validation of such systems. An independent validation is often a challenge. Different explainability methods exist based on different principles.

- Model-agnostic methods: LIME and SHAP.
- Model-specific methods: DeepLIFT and Integrated Gradients.
- Interactive methods: Explainable Boosting Machine (EBM) and What-If Tool.

In our proposed approach, before we use an external system for explanations, we first model the steering parameters and design a prompt to provide us with the desired explanations. This is usually done to fine-tune the structure of your results generated by an LLM.

4 Experiments

The design-time compliance, informed by *ante hoc* explanations, is in general out of scope of our experimental setup. However, it can be noted, that such compliance can be assured by reference to the datasets used, and training procedures the model was used with. This refers to many phases of GenAI model development, including its pre-training, fine-tuning for a specific tasks, and human-assisted training down the line. This can be realised for art. 13 EU AU Act compliance.

The focal point of our experimental setup is art. 14, as it relates *post-hoc* explainability for human oversight. The structure of our system that realises

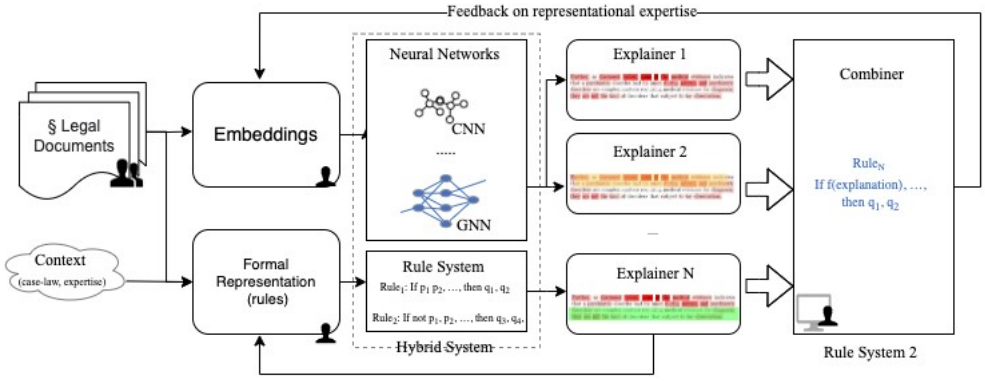


Figure 2: Overview of implemented system

this ascpect of transparency is provided in fig. 2. It sports a pluggable architecture, whereas the centre, classification module, can accomodate any classifier. In experiments presented in this paper, we have used an LLM, with prompt engineered for a classification exercise. Second part of the system allows for a pluggable set of explainers to be introduced. This enables a contrastive approach, i.e. the comparison of explanations offered by various methods. This allows for:

- Contrasting different classifiers of a larger system to assure they all are sound regarding their rationale for a decision.
- Making a comparison between various explainers, to align their predictions by hyperparameter tuning or spotting faulty ones.
- Combinging explanations to give a broader picture of a system's inner working, eventually producing a more holistic understanding.

As for the concrete application, in our earlier paper we have shown that it is in principle possible to use well-known explanation generation methods to create ones for LLM system [3]. Therein, we have compared both *internal* and *external* explanation. For internal, we have instructed the model to output the features that are decisive for its prediction; for external, a well-known SHAP method has been applied. SHAP uses a game-theoretic concept of Shapley values and perturbations of base sample to identify the most important features for a model.

Our already-implemented system uses an LLM (we have used LLAMA-2-13b, Vicuna-1.5-13B and GPT-4, with the best results achieved with the last one) for a classification task. No fine-tuning was performed and we relied on prompt engineering to create a classifier. For the dataset, we have chosen a fictitious domain, a well-known hospital welfare benefits dataset [1]. Using a fictious domain frees us from having to do with fuzziness and mistakes present in real

world datasets, and allows to focus on the domain at hand. In this dataset, we are to grant a benefit for a person for their spouse's hospital visit. In the original dataset, the approval of such benefit is dependant on 12 variables, concerning the facts such as the payment of national health system's contributions and being of relevant age. For the sake of the experiment presented herein, we will simplify the domain and rely only on three features: whether the person applying for the benefit is devoid of sufficient savings (`capital <= 3000`), is a spouse of a person in the hospital (`is_spouse = true` and is not absent in the country (`is_absent = false`). Limited domain makes the presentation easier and significantly reduces the computational complexity of the problem. The SHAP we used for external explanations, a perturbation-based explainer is very resource intensive. We have not yet implemented the distributed calculations for the SHAP explainer [2], thus for this reason, as well as the presentation simplicity, we have decided to limit the domain. We clearly state this is a limitation of our study, when LLMs are chosen as the classifiers.

5 Results

In our case, there are three rules that have to be fulfilled for the benefit to be granted, and each rule pertains to a single feature from the test set. We have created a system that uses two explainers that can be used by a system deployer to test its inner-wokring, they can be also passed in some form to the end-user as a part of a decision's justification. In addition to internal and external explanation, in the case of our system, ground truth can also be used by explanation combiner. Establishing effective governance and transparency for productionized applications is one of the core requirements of the European Union's Artificial Intelligence Act. We adapted a well-known notion called Guardrails to our framework. Guardrails provide the tools for implementing a set of safety controls that monitor and dictate a user's interaction with an LLM application. They are a set of programmable, rule-based systems that sit between users and foundational models to ensure that the AI model is operating within the defined principles of an organization.

In our proof-of-concept implementation, we are able to compare and contrast all three explanations. This can be done in principle by a Python script, or any sufficiently sophisticated rule engine, like Drools, or with the use of controlled natural language-based system (CNL). CNLs are a subset of natural language that restrict its grammar and vocabulary to reduce its vagueness, eventually allowing for automatic down the line processing. They have already been used in the business, thus this shows how our system can be further integrated into existing infrastructure. Obviously, it is also possible to employ yet another downstream LLM for comparison of explanations.

R1	R2	R3	GT	GPT4 Pred.	GPT4 Features	SHAP_GPT4	Jaccard score
F	F	F	T	T	F1, F2	F3, F1	1/3
F	T	F	F	F	F2, F3	F2, F3	1
T	T	F	F	F	F1	F1, F2	1/2
F	F	F	T	F	F3	F3, F2	1/2
F	T	T	F	F	F2, F3	F3, F1	1/3
F	F	T	F	F	F3	F3, F1	1/2
T	T	T	F	F	F1, F3	F3, F1	1
T	F	F	F	T	F2, F3	F3, F1	1/3

Table 1: Different types of samples assessed by GPT-4, and comparison of explanations given by SHAP (SHAP_GPT4 column, showing 2 most important features) and the model itself (GPT4 features column). Jaccard Score column denotes how many top-2 features were identified as the same by two explainers T denotes True, F - False, R{N} is for Rules Violated, GT = Ground Truth, F{N} is for Features where F1 = 'is_spouse', F2 = 'is_absent' and F3 = 'capital_resources'.
In our streamlined case, feature F{N} applies to rule R{N}, that is - each rule makes use of only a single feature. The table's results are adapted from [3]

For the experiment presented herein, we have used a Jaccard Score to compare the two most important features as identified by SHAP and GPT-4. In our case, the mean score for all the samples is .56, which shows there is a signifacant disagreement between internal and external explanations. This can a signal for the human performing oversight to look into the model's - or explainers - performance.

6 Conclusion

In this paper, we have presented the conceptual framework for business-oriented implementation of explainable models. We have presented how this framework can be aligned to draft of EU AI Act, as well as presented a proof-of-concept system implementing this framework. Our experimental results are based on our earlier study (regarding the explanations generation). Herein, we have shown that different explanations can be contrasted and compared.

7 Acknowledgements

This research is supported by Notre Dame-IBM Technology Ethics Lab, under the project *Domain-specific explainable artificial intelligence for AI auditing*, grants 262812LG and 262812SR, as well as the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw, under grant G93-1608. Such support does not constitute endorsement by the sponsor of the views expressed in this publication.

References

[1] Trevor Bench-Capon. Neural networks and open texture. In *Proceedings of the 4th international conference on Artificial intelligence and law*, pages 292–297, 1993.

[2] Roie Chen. Distribute shap calculations, Jul 2023.

[3] G&#243, Łukasz Rski, and Shashishekar Ramakrishna. Challenges in Adapting LLMs for Transparency: Complying with Art. 14 EU AI Act. In *Legal Knowledge and Information Systems*, pages 275–280. IOS Press, 2023.

[4] Guido Governatori, Mustafa Hashmi, Ho-Pun Lam, Serena Villata, and Monica Palmirani. Semantic Business Process Regulatory Compliance Checking Using LegalRuleML. In Eva Blomqvist, Paolo Ciancarini, Francesco Poggi, and Fabio Vitali, editors, *Knowledge Engineering and Knowledge Management*, volume 10024, pages 746–761. Springer International Publishing, Cham, 2016. Series Title: Lecture Notes in Computer Science.

[5] Balint Gyevnar, Nick Ferguson, and Burkhard Schafer. Bridging the transparency gap: What can explainable ai learn from the ai act?, 2023.

[6] Philipp Hacker. Ai regulation in europe: From the ai act to future regulatory challenges. *arXiv preprint arXiv:2310.04072*, 2023.

[7] Mustafa Hashmi, Guido Governatori, Ho-Pun Lam, and Moe Thandar Wynn. Are we done with business process compliance: state of the art and challenges ahead. *Knowledge and Information Systems*, 57(1):79–133, October 2018.

[8] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.