Deep Natural Language Feature Learning for Interpretable Prediction

Felipe Urrutia[1, 2], Cristian Buc[1], and Valentin Barriere[1, 2]

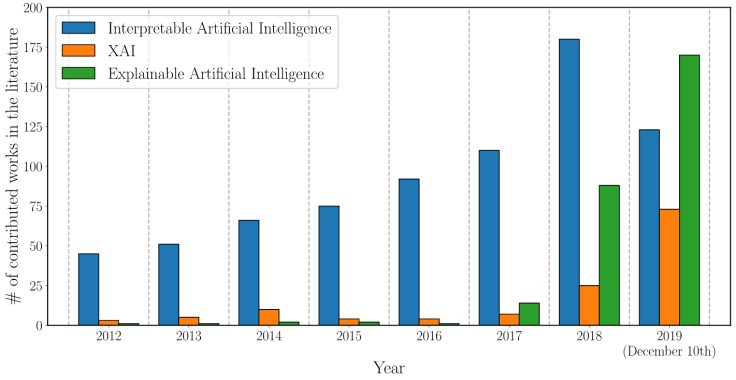
December 29, 2023

[1]Centro Nacional de Inteligencia Artificial, Macul, Chile
[2]Department of Computer Science, Universidad de Chile, Santiago, Chile

Introduction and Related Work

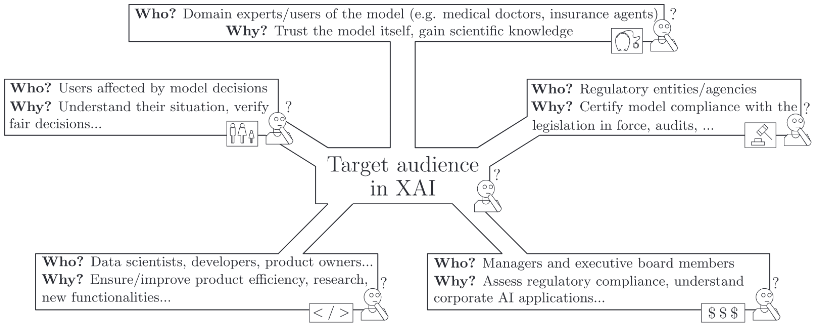
The Challenge of Explainable AI (XAI) in Deep Learning

- The increasing use of AI models in healthcare (Norgeot et al., 2019), justice (Dass et al., 2022), and finances (Heaton et al., 2017)
- Deep Learning models excel at complex decision-making but are often seen as "black boxes" (Castelvecchi, 2016)
- Regulations and growing interest in Explainable AI (XAI) have emerged (Goodman and Flaxman, 2017; Russell et al., 2015)



The Challenge of Explainable AI (XAI) in Deep Learning

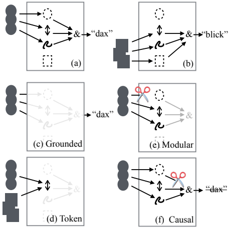
- XAI aims to provide human-interpretable information on model behavior (Gunning et al., 2019; Arrieta et al., 2020)



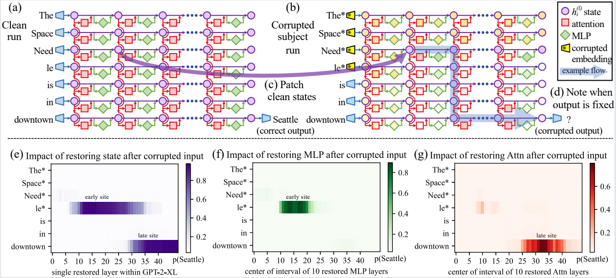
- In the NLP domain, XAI can be divided into **representational** and **practical** categories

Representational

- Focuses on grasping the underlying structure of representations
- Transformer-based architectures develop abstract symbolic or compositional representations (Lovering and Pavlick, 2022; Li et al., 2022b)



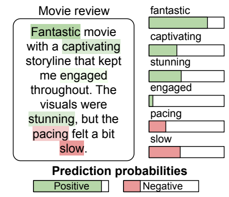
- Sparse representations of conceptual knowledge can be located and edited to induce different predictions (Meng et al., 2022)



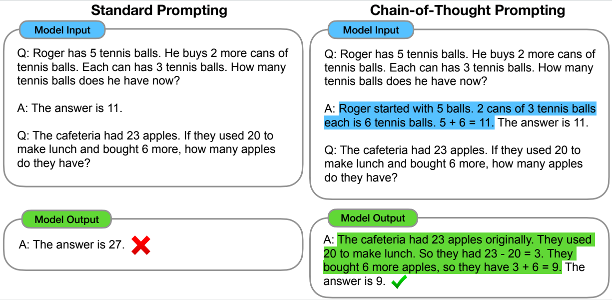
Explainable Deep Learning in NLP

Practical

- Analyzes model outputs and enhancing explainability
- Analyzing model behavior when perturbing inputs (Tulio Ribeiro et al., 2016; Fel et al., 2023; Lundberg and Lee, 2017)

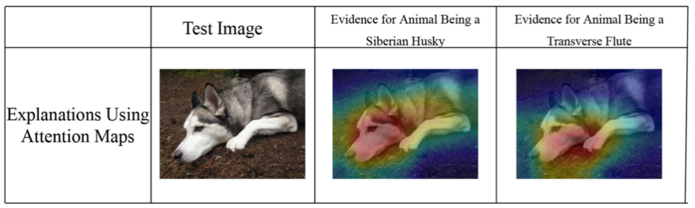


- Prompting to increase explainability, such as Chain-of-Thought (Wei et al., 2022; Wang et al., 2022; Zhao et al., 2023; Lyu et al., 2023)



The Shift Towards Interpretability

- Recent push to focus on interpretable models for high-stakes decisions (Rudin, 2019)



- Motivated by significant errors made by black-box models in real-life situations (Angwin et al.; Wexler, 2017)
- Differentiating between *explainable black-box* and *interpretable white-box* models (Rudin, 2019)

The Shift Towards Interpretability

Explainability

Relies on algorithms aiming to **explain** the model predictions

Interpretability

Relies on the possibility to **know exactly why** the model is making a prediction

- Challenges in achieving true interpretability, e.g., limitations of methods like CoT to produce externalized reasoning (Radhakrishnan et al., 2023)

Natural Language Learned Features

Motivation

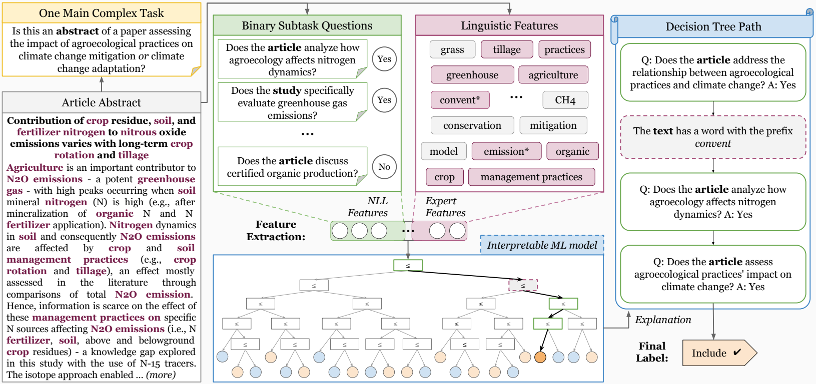
- **Context:** Interpretability in critical decisions to prevent real-life errors from black-box models
- **Aim:** Reconcile black-box LLM and interpretable ML models
- Combine LLM's strength (e.g. zero-shot abilities) and ML classifier's explainability

Our Approach

- Leverage LLM to decompose complex tasks into simpler sub-tasks
- Utilize sub-tasks with a medium-sized language model for interpretable features
- Improve ML classifiers, e.g., Decision Trees with readable decision paths

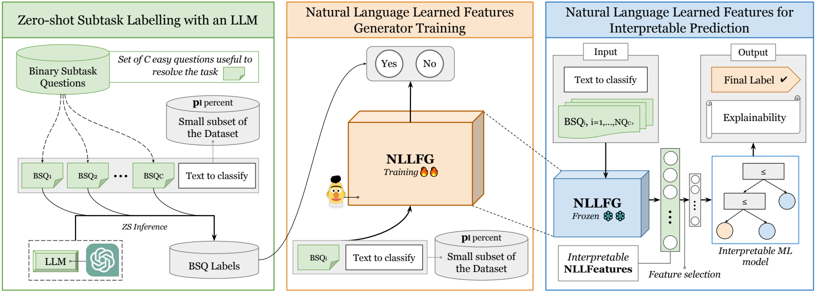
Methodology

Overview of the proposed system



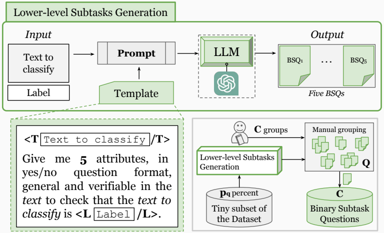
Extraction of Natural Language Learned Features and Expert Features in order to understand the decision process of an interpretable model for complex task solving

A Method for Natural Language Learned Features

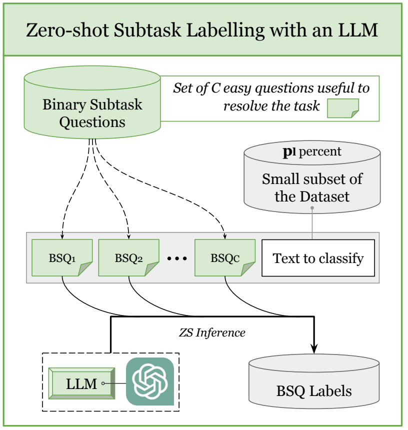


Full process of subtask labelisation, NLLF-generator training, NLLF generation and integration

Zero-shot Subtask Labelling with an LLM



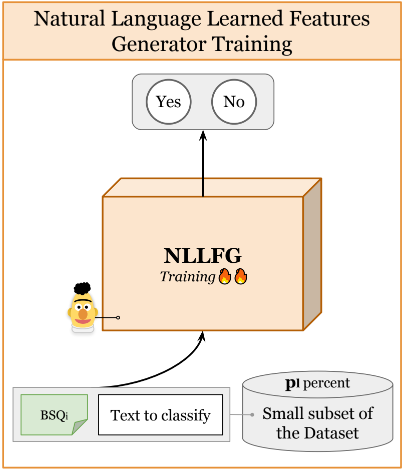
- LLM to decompose complex tasks into simpler sub-tasks as Binary Subtask Questions (BSQs)
- LLM labels each sub-task in zero-shot format on a small subset of the dataset



NLLF Generator (NLLFG) Training

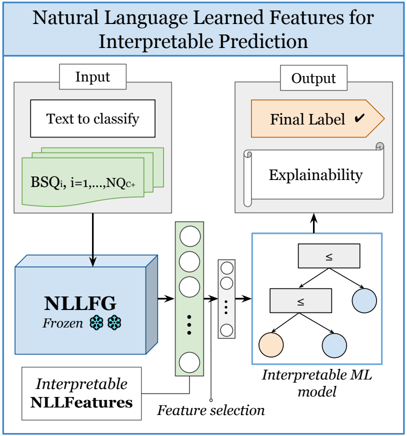
- Sub-tasks with a unique medium-sized LM
- LM trained for all the BSQs using the labels generated by the LLM over the small subset of the dataset
- Interpretable features are the logits of the LM
- NLI-like training:

[CLS] Text [SEP] BSQ [SEP]



NLLF for Interpretable Prediction

- NLLFG can take any question formulated in natural language
- NLLF as input for an Decision Tree
- Explanations provided in the form of a decision path in the tree
- The selected NLLF can be used for any other model

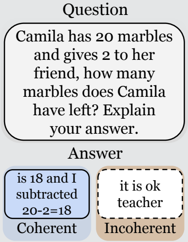


Experimental Setup

Datasets

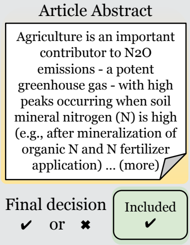
Incoherent Answer Detection

- Data from the Chilean e-learning platform ConectaIdeas
- 15,435 answers to 700+ different open-ended math questions
- The answers' (in)coherence were manually annotated by several teachers



Scientific Abstract Classification

- 15,000+ articles from Web of Science database
- 1,983 relevant articles for systematic literature review on Agroecology and Climate Change
- Articles tagged by two annotators for relevance, with a third annotator arbitrated in 14% of cases

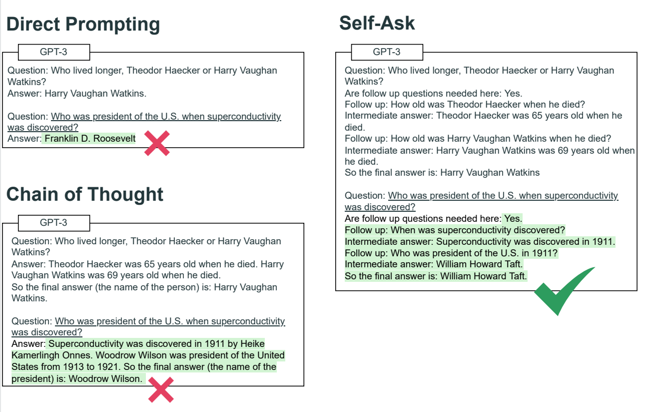


Baselines

- **In-Context-Learning with LLM**: ChatGPT in 0- and 4-shot format using three prompt variants:

Prompt variants

- **Vanilla** (Brown et al., 2020)
- **Chain-of-thought** (Wei et al., 2022)
- **Self-ask** (Press et al., 2023)



Baselines

- **Black-box neural networks**: Transformer-based:
  - **BERT** in English (Devlin et al., 2018), and
  - **BETO** in Spanish (Cañete et al., 2020)
- **Interpretable models**: Decision Tree using Bag-of-Words
- **Other Features**
  - **Expert Features:** Linguistics Features made by expert
  - **Natural Language Learned Features** (NLLF)

Results

Model	Variant	Params	Explainability	F1-score	
				IAD	SAC
ChatGPT	0-shot	$\sim 10^{11}$	✗	31.33	35.23
	4-shots		✗	38.92	38.92
	0-shots CoT		✓	36.33	41.59
	4-shots CoT		✓	56.45	<u>62.72</u>
	0-shots SA		✓	33.84	48.13
	4-shots SA		✓	<u>62.17</u>	59.50
BERT	Vanilla	$\sim 10^{8}$	✗	67.08	67.72
	EF		✗	75.10	66.90
	NLLF		✗	71.48	68.75
	NLLF+EF		✗	<u>76.45</u>	<u>73.63</u>
Decision Tree	BoNG	$\sim 10^{2}$	✓	14.97	65.15
	EF		✓	73.77	64.95
	NLLF		✓	55.56	62.25
	NLLF+EF		✓	<u>78.09</u>	<u>67.75</u>
	NLLF+BoNG		✓	60.47	66.20
	NLLF+EF+BoNG		✓	<u>78.09</u>	67.41

F1-score of all the configurations and models for Incoherent Answer Detection (IAD); and (Macro) F1-score for the Scientific Abstract Classification (SAC). Using Expert Features (EF), NLLF, and Bag-of-N-Grams (BoNG).

Conclusion and Future Works

Conclusion

Contributions

- Method for breaking down complex tasks into easier sub-tasks using natural language binary questions
- Generation of Natural Language Learned Features (NLLF)
- Enhancement of any classifier using NLLF
- Use of NLLF as input for an easy-to-interpret machine learning model like a decision tree

Future Directions

- Investigate real-world applications, especially in education
- Understand practitioners' preferences for explainability
- Explore scalability to more complex tasks

Thanks for listening!