

# Why Is This Green? LLM-Based Explanations of Implicit Green Practices in Social Media

Anna Glazkova<sup>1,2</sup>, Olga Zakharova<sup>1,2</sup>, Daria Lebedeva<sup>3</sup>

<sup>1</sup> Carbon Measurement Test Area in Tyumen' Region (FEWZ-2024-0016), Tyumen, Russia

<sup>2</sup> University of Tyumen, Tyumen, Russia

<sup>3</sup> National Research University Higher School of Economics, Moscow, Russia  
a.v.glazkova@utmn.ru, o.v.zakharova@utmn.ru, dlebedeva@hse.ru

## Abstract

Identifying green practices in social media is not merely a matter of lexical matching. Many green practices are expressed implicitly, rely on shared background knowledge, or are embedded in broader contextual narratives. In this paper, we investigate how large language models (LLMs) explain expert annotations of green waste management practices and how they rationalize classification errors made by a fine-tuned model (mBART) on a Russian social media corpus (GreenRu). We analyze explanations generated by two LLMs (T-lite and GigaChat) in two settings: (1) explaining gold expert-assigned labels and (2) interpreting erroneous model predictions. Our qualitative and micro-quantitative analysis shows that green practices are frequently inferred through contextual reasoning rather than explicit terminology. Error patterns of mBART reveal overgeneralization, associative misinterpretation (e.g., linking food sharing to waste recycling), and detection of practices where none are present. We further compare explanatory strategies of the two LLMs. T-lite tends to rely on lexical cues and surface markers that may create an impression of a practice, while GigaChat more often reconstructs broader contextual interpretations. Expert feedback highlights limitations of formal textual analysis, sensitivity to missing contextual knowledge, and difficulties in aligning model reasoning with expert conceptual boundaries. Our findings suggest that explanation-based analysis is a productive tool for diagnosing classification errors and refining annotation guidelines. More broadly, the study demonstrates that modeling implicit sustainability discourse requires contextual grounding and deeper semantic integration beyond keyword-based approaches.

**Keywords:** green practices, social media, implicit information extraction, large language models, explainability, error analysis, environmental NLP

## 1. Introduction

The growing societal attention to environmental sustainability has led to the rapid expansion of online communities centered around ecological lifestyles, responsible consumption, and waste management. Social media platforms serve not only as spaces for information exchange but also as arenas where green practices are negotiated, normalized, and disseminated. Detecting such practices automatically is essential for large-scale analysis of grassroots environmental engagement, public discourse monitoring, and policy-oriented research.

However, identifying green practices in user-generated content is not a trivial text classification problem. Many environmentally relevant actions are expressed implicitly, rely on shared background knowledge, or are embedded in broader narratives about everyday life. For instance, bringing items to a named local initiative may constitute a practice of sharing or reuse, even if no explicit sustainability terminology is used. This makes green practice detection fundamentally a task of contextual interpretation rather than simple lexical matching.

Previous work in environmental natural language processing (NLP) has addressed related tasks such as climate change discourse analysis, environ-

mental stance detection, and sustainability-related information extraction (Beckles and Heidke, 2025; Peura et al., 2025; Zhang et al., 2024; Longo and Longo, 2025; Barz et al., 2025; Adam and Kogler, 2025). However, most studies focus either on topic modeling, sentiment and stance analysis, or explicit mention detection, rather than fine-grained behavioral practice identification—especially when practices are only indirectly expressed.

From a methodological perspective, our study intersects with research on multi-label text classification and explainability. Multi-label classification has been widely applied in various NLP domains. For example, Kementchedjnieva and Chalkidis (2023) demonstrated the effectiveness of encoder–decoder architectures for multi-label classification in biomedical and legal domains. Similarly, multilingual sequence-to-sequence models such as mBART (Tang et al., 2021) have proven competitive for complex classification tasks when reformulated as text generation.

Beyond classification performance, recent years have seen growing interest in explanation-based analysis of model behavior. Explainable AI in NLP has evolved from feature attribution methods (e.g., attention visualization, gradient-based saliency) toward natural language explanations generated by

LLMs (Zhao et al., 2024). Instruction-tuned LLMs can provide post-hoc rationalizations of predictions, enabling qualitative diagnosis of systematic errors (Brown et al., 2020; Kamoi et al., 2024). Such approaches are increasingly used not only to justify model outputs but also to probe implicit reasoning patterns and uncover annotation inconsistencies.

This work also relates to research on implicit information extraction and pragmatic inference in NLP. Detecting implicitly expressed events, relations, or intentions remains a longstanding challenge. Prior studies in implicit sentiment analysis, implicit hate speech detection, and implied event extraction show that models often rely on surface lexical associations rather than deeper semantic integration (ElSherief et al., 2021; Zhu et al., 2024; Loukachevitch et al., 2025; Masud et al., 2025). The difficulty increases in domain-specific contexts, where shared community knowledge plays a crucial role (Galli et al., 2025; Sengupta et al., 2025).

Within sustainability studies, the conceptualization of green practices as socially embedded behaviors rather than isolated keywords has been discussed in social science research (Zakharova et al., 2022). The GreenRu dataset (Zakharova and Glazkova, 2024) operationalizes this perspective by annotating concrete waste management practices at the sentence level in Russian-language social media posts. Unlike broader thematic datasets, GreenRu focuses on specific behavioral categories (e.g., waste sorting, sharing, refusing purchases), enabling fine-grained multi-label modeling.

In this paper, we move beyond standard performance evaluation and investigate how LLMs explain both correct expert annotations and systematic classification errors produced by a fine-tuned mBART classifier. We compare two Russian-language instruction-tuned LLMs: T-lite and GigaChat. Our analysis addresses three research questions:

1. How do LLMs reconstruct the reasoning behind expert-assigned green practice labels, especially when practices are expressed implicitly?
2. How do LLMs rationalize systematic classification errors, and what types of textual cues do they identify as misleading?
3. To what extent do explanation patterns reveal mismatches between expert conceptual boundaries and model-internal associations?

Our findings demonstrate that explanation-based analysis is a productive diagnostic tool for sustainability-oriented NLP tasks. Green practice detection emerges as a problem of contextual grounding and shared knowledge modeling rather than mere lexical classification. Moreover, comparing LLM-generated rationales with expert reflections

reveals not only model weaknesses but also latent ambiguities in annotation guidelines.

More broadly, the study contributes to the emerging intersection of environmental NLP and explainable AI by showing that modeling implicit sustainability discourse requires deeper semantic integration and explicit attention to contextual knowledge structures.

The remainder of the paper is structured as follows. Section 2 describes the dataset and task formulation, Section 3 presents the experimental setup, Section 4 reports the results, and Section 5 discusses implications of LLM-based explanations.

## 2. Dataset and Task

### 2.1. GreenRu Dataset

This study employs the GreenRu dataset (Zakharova and Glazkova, 2024)<sup>1</sup> to identify references to green waste management practices in Russian-language social media content. The dataset comprises 1,326 Russian-language posts, averaging 880 characters in length, sourced from online environmentally conscious communities.

The annotation of GreenRu was conducted at the sentence level using a multi-label scheme to capture the specific green waste practices mentioned, with sentences averaging 110 characters. The annotation framework is based on nine categories of green waste practices (Zakharova et al., 2022):

1. *waste sorting*, referring to the separation of waste by material type;
2. *studying product labeling* to identify packaging as a waste category;
3. *waste recycling*, the process of converting waste into reusable materials;
4. *signing petitions* as a means of influencing policy;
5. *refusing purchases* to minimize consumption and environmental impact;
6. *exchanging* unwanted items or services for desired ones;
7. *sharing* goods with others, either free or for a fee;
8. *participating in initiatives promoting responsible consumption*, such as workshops, festivals, or educational events;
9. *repairing* items instead of discarding them.

---

<sup>1</sup><https://github.com/green-solutions-lab/GreenRu>

The mentions of green practices are unevenly distributed. The most common practice is waste sorting (1,835 mentions), while the rarest is repairing (13 mentions).

## 2.2. Green Practice Detection

In this study, we explore the task of detecting mentions of green waste practices in social media. By green waste practices, we refer to concrete actions, behavioral patterns, or organizational decisions aimed at reducing environmental impact, conserving resources, promoting circular economy principles, or supporting broader environmental sustainability goals.

We formulate this problem as a multi-label text classification task. Given a textual fragment (typically a sentence or a short passage), the model must determine whether it mentions green practices.

## 3. Experimental Setup

### 3.1. Classification Model

For multi-label text classification, we used mBART<sup>2</sup> (Tang et al., 2021), a multilingual sequence-to-sequence transformer architecture introduced by Facebook AI Research. In previous studies (Glazkova and Zakharova, 2025), mBART achieved the best overall performance on green practice detection, outperforming both traditional machine learning approaches (e.g., kNN trained on sentence embeddings) and transformer-based models including BERT-based classifiers. A similar effectiveness of encoder-decoder models for multi-label text classification was demonstrated in the work of Kementchedjhieva and Chalkidis (2023) for legal and biomedical texts.

Following the paper of Glazkova and Zakharova (2025), we fine-tuned mBART for 20 epochs using a maximum sequence length of 256 tokens. The model input consisted of sentences from the GreenRu dataset, and the output was a comma-separated list of the practices mentioned in the corresponding sentence.

### 3.2. LLMs for Explanation

For generating explanations, we used two instruction-based LLMs:

- T-lite<sup>3</sup> (Stoianov et al., 2025), a Russian-language model built upon the Qwen 3 architecture, 8B params.

---

<sup>2</sup><https://huggingface.co/facebook/mbart-large-50>

<sup>3</sup><https://huggingface.co/t-tech/T-lite-it-2.1>

- GigaChat<sup>4</sup> (Mamedov et al., 2025), a Russian-language model built on a Mixture-of-Experts (MoE) architecture with 10B total and 1.8B active parameters, it incorporates Multi-head Latent Attention (MLA) and Multi-Token Prediction (MTP) to achieve high inference throughput.

### 3.3. Explanation Settings

To evaluate the explanations generated by LLMs, we constructed two subsets of entries from the test portion of the GreenRu dataset. The dataset itself exhibits a naturally imbalanced distribution of practice types, reflecting the real-world prevalence of different green practices rather than an artificially balanced annotation scheme. Each subset contained 100 entries selected through randomized sampling while ensuring that all practice types were represented. Full class balance, however, could not be achieved due to the multi-label nature of the dataset: rare practices frequently co-occur with more common ones, making strict balancing infeasible without distorting the original data distribution.

The first subset (*random*) consisted of randomly selected entries together with their gold annotations of green practice mentions. The second subset (*error-only*) included entries for which none of the labels predicted by the mBART model matched the expert annotations. For these entries, we collected both the model predictions and the corresponding gold annotations. Examples of the selected entries and typical mBART errors are presented in Table 1. The normalized frequencies of gold-labeled practices in both subsets are shown in Figure 1. The figure demonstrates that while the overall distributions remain dominated by more frequent practices, both subsets preserve coverage of the full practice inventory, allowing a qualitative comparison of explanation behavior without substantially altering the underlying data distribution.

The prompts for analyzing expert annotations and mBART errors are presented in Table 2. The first prompt was intended to understand how human experts interpret implicit or explicit mentions of green practices, revealing the linguistic cues and contextual reasoning behind the gold labels (RQ1). The second prompt was designed to trace mBART errors back to specific textual elements, helping us identify whether mBART was misled by ambiguous wording, false positives (i.e. cases where the model erroneously detects a practice), or superficial patterns in the data (RQ2).

For both models, we used a generation temperature of 0.5. This value was chosen as a compromise between determinism and variability in generation. In preliminary experiments, lower values resulted

---

<sup>4</sup><https://huggingface.co/ai-sage/GigaChat3-10B-A1.8B-bf16>

Text	Annotated practices (gold)	Annotated practices (mBART)
Упаковка с такой маркировкой не может быть переработана в принципе и закончит свой жизненный цикл на мусорном полигоне или в печи мусоросжигательного завода. { <i>Packaging with this labeling cannot be recycled at all and will end its life cycle in a landfill or in the furnace of an incineration plant</i> }	Изучение маркировки товаров { <i>Studying product labeling</i> }	Переработка отходов { <i>Waste recycling</i> }
В Великобритании изобрели #устройство для идентификации отходов - Recycling Identifying Device (R.I.D.). { <i>In the UK, a waste identification device called the Recycling Identifying Device (R.I.D.) has been invented</i> }	Сортировка отходов { <i>Waste sorting</i> }	Изучение маркировки товаров { <i>Studying product labeling</i> }
Фудшеринг - проект по спасению еды от утилизации, направленный на сохранение ресурсов нашей планеты и во имя борьбы с перепотреблением. { <i>Food sharing is a project aimed at saving food from disposal, preserving our planet's resources, and combating overconsumption</i> }	Обмен { <i>Sharing</i> }	Переработка отходов, отказ от покупок { <i>Waste recycling, refusing purchases</i> }

Table 1: Examples of entries and errors.

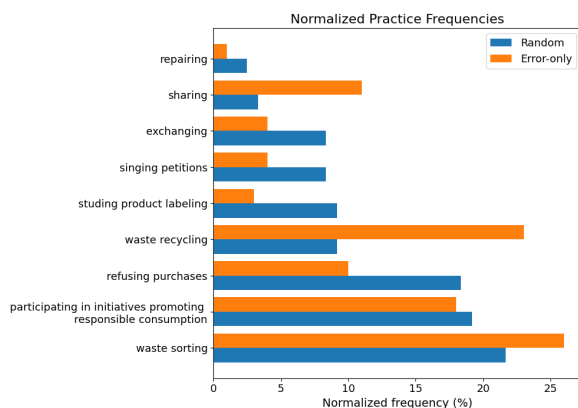


Figure 1: Normalized practice frequencies in the selected subsets.

in more repetitive explanations, while higher values increased variability without improving interpretability. The maximum number of tokens for explaining the expert annotations was set to 512, and for explaining the mBART annotations, it was set to 1024 tokens. In the second case, the maximum number of tokens was larger because the LLMs tended to generate longer responses to the second prompt. The actual energy consumption and carbon emissions of LLMs for 100 entries are given in Table 3.

#### 4. Results

The average explanation length for the random subset was 1,346.67 characters for T-lite and 442.96 characters for GigaChat. The explanation length for gold labels in the error-only subset was 1,275.84 (T-lite) and 375.07 (GigaChat). The explanation length for mBART errors was 2,979.57 (T-lite) and 1,861.24 (GigaChat). On average, T-lite produces

longer and more lexically focused explanations, while GigaChat generates shorter, more abstractive justifications.

In the majority of cases, both LLMs explicitly refer to lexical units or short phrases from the input text as justification. However, T-lite more frequently relies on isolated lexical triggers, whereas GigaChat more often reconstructs larger semantic fragments (see Table 4). Both models often explicitly indicate whether a practice is expressed directly or implicitly. GigaChat more consistently signals implicitness (e.g., “the practice is not directly stated but inferred from...”), while T-lite more often assumes explicitness even when expert reasoning is contextual. In analyzing the errors of mBART, both models are generally able to identify plausible textual cues that could have triggered the incorrect prediction (RQ2). This suggests that many classification errors stem from systematic lexical or semantic associations rather than random noise.

The explanations produced by both LLMs were additionally analyzed in order to assess the reasoning strategies underlying their decisions. For each explanation column, we manually evaluated two indicators: (i) whether the model supported its decision with explicit lexical evidence by referring to fragments or linguistic cues present in the source text, and (ii) whether the model explicitly indicated that the described practice was expressed implicitly rather than directly stated. The analysis demonstrated that both models predominantly relied on textual grounding: lexical evidence was used in more than 95% of explanations across models. However, substantial differences were observed in how the models handled implicit meaning. The T-lite model identified practices as implicitly expressed in approximately 84% of cases,

Prompt ID	Prompt text
1	<p>The following sentence was annotated by experts as describing the following green practices: [PRACTICE_NAME] Sentence: "[TEXT]"</p> <p>Explain how exactly these practices are expressed in the text. If they are expressed implicitly, indicate what this interpretation is based on. If there is no explicit expression of these practices in the text, state this.</p>
2	<p>The model interpreted the following sentence as describing the following green practices: [PRACTICE_NAME] Sentence: "[TEXT]"</p> <p>Explain which elements of the text could have created the impression that these practices are being referred to. If this impression is weak or misleading, explain why.</p>

Table 2: Prompts used for generation. The prompts are given in English (translated from Russian). In cases where an entry had a single green practice annotated by experts or mBART, the corresponding singular phrasing was used.

Model	Prompt 1			Prompt 2		
	Time	Energy (kWh)	CO <sub>2</sub> eq (g)	Time	Energy (kWh)	CO <sub>2</sub> eq (g)
T-lite (8B)	0:28:45	0.1036	51.69	1:08:24	0.2425	120.93
GigaChat (1.8B active)	0:08:57	0.0287	7.28	0:40:35	0.1246	31.56

Table 3: The actual energy consumption and carbon emissions of LLMs for 100 entries using Google Colab Pro (NVIDIA A100-SXM4-40GB). The values are obtained with Carbontracker (Anthony et al., 2020).

whereas GigaChat did so considerably less frequently, in about 27% of explanations. In the analysis of model error interpretations, plausible textual triggers motivating incorrect predictions were identified in the clear majority of cases (over 90%), with only a small number of explanations lacking such triggers. These findings indicate that while both models ground their reasoning in the source text, they differ markedly in their tendency to infer implicit practices.

A central finding of our experiment is that green practice detection is fundamentally a task of contextual interpretation rather than lexical matching (RQ1). Many practices in GreenRu are expressed implicitly. For example, shared use initiatives (e.g., donation-based stores or exchange platforms) are not always described through explicit sustainability vocabulary. Human experts rely on background knowledge: they understand the institutional purpose of specific initiatives and interpret actions (e.g., bringing items to a named store) as part of a broader green practice. LLMs demonstrate partial sensitivity to this phenomenon. When contextual knowledge is available or widely encoded (e.g., widely known eco-initiatives), explanations become more accurate and coherent. When background knowledge is missing, models revert to surface-level lexical interpretation. In several cases, models explicitly mention “lack of context” or “insufficient

information” as a limitation. This meta-commentary indicates that LLMs can recognize epistemic uncertainty, yet they still tend to produce a forced interpretation.

Analyzing explanations for incorrect mBART predictions reveals systematic error types (RQ2). First, the model predicts a broader practice category instead of a more specific one (e.g., interpreting “bringing recyclables” as “recycling” rather than “waste sorting”). Second, closely related practices are confused due to associative links (e.g., food sharing interpreted as waste recycling because of lexical proximity between “food” and “waste”). Third, some keywords (e.g., “free,” “eco,” “volunteers”) activate sustainability interpretations even when the expert label depends on a more specific conceptual framing. Finally, the experts distinguish stages within a sustainability chain (label reading → sorting → transport → recycling), whereas the model treats them as a single undifferentiated process.

## 5. Expert Assessment of LLM Explanations

We invited green waste practice experts who participated in the original annotation process to review LLM-generated explanations. The expert assessment was conducted in a two-stage procedure.

<b>Text</b>	Вся макулатура будет переработана для вторичного использования. <i>{All waste paper will be recycled for reuse}</i>
<b>Gold labels</b>	Переработка отходов <i>{Waste recycling}</i>
<b>mBART labels</b>	Сортировка отходов <i>{Waste sorting}</i>
<b>Explanations for the gold labels (Prompt 1)</b>	
T-lite	<p>The practice of waste recycling is explicitly expressed in the sentence: "All waste paper will be recycled for reuse."  Explanation:  The text directly states: "will be recycled" — this is a verb denoting the action of recycling.  It specifies what exactly is being recycled — waste paper (a type of waste based on paper and cardboard).  It also indicates the purpose of the action — reuse, which aligns with the environmental goal of recycling: preventing waste from ending up in landfills, conserving resources, and reducing the ecological footprint.  Conclusion:  The practice is explicitly expressed through the verb "will be recycled" and the specification of the object (waste paper) and the purpose (reuse).  Thus, the sentence fully corresponds to the description of the environmental practice of waste recycling.</p>
GigaChat	<p>In this sentence, the green practice of 'waste recycling' is explicitly expressed through the use of the verb 'will be recycled.'  This indicates that all waste paper will undergo the recycling process for the purpose of reuse.</p>
<b>Explanations for the mBART labels (Prompt 2)</b>	
T-lite	<p>Excellent question. Let's examine why the sentence "All waste paper will be recycled for reuse" could be interpreted as describing an environmental practice, specifically waste sorting, and to what extent this impression is justified.  Elements of the text that created the impression of "waste sorting":  1. The mention of "waste paper"  — Waste paper consists of paper and cardboard waste that must be sorted during the collection and recycling stages.  — The word "waste paper" itself indicates the separation of waste by type, i.e., sorting.  2. The phrase "will be recycled"  — Waste recycling always begins with sorting. Without sorting materials by type (paper, plastic, glass, etc.), recycling is impossible or extremely inefficient.  — This creates an association with a logical sequence: first sorting → then recycling.  3. The goal — "reuse"  — This is a key goal of green practice aimed at resource conservation.  — Sorting is one of the main ways to ensure the reuse of waste, especially waste paper.  Conclusion: the impression of "waste sorting" is justified and strong  Despite the fact that the sentence does not use the word "sorting", all the key elements point to the practice of waste separation:  - waste paper → a type of waste that requires sorting  - recycling → requires preliminary sorting  - reuse → a goal achieved through sorting  Therefore, the model's interpretation as "a description of an environmental practice — waste sorting" is justified and logical.  Why does it not mislead?  - There are no contradictions between the text and the interpretation.  - All terms and logic correspond to real-world waste management practices.  - Even if the emphasis is on recycling, it cannot be effective without sorting, especially for waste paper.  Summary:  Yes, the impression of "waste sorting" is not weak, but quite justified.  The sentence clearly indicates an environmental practice related to sorting and does not mislead — on the contrary, it reflects the standard and correct approach to recycling waste paper.  If the goal were, for example, "waste incineration" or "landfilling," the interpretation would be different. But here — everything points to an environmental practice with an emphasis on sorting and reuse.</p>
GigaChat	<p>The elements of the text that could create the impression that it refers to waste sorting:  The word "waste paper" — This word is directly associated with the process of recycling paper and cardboard, which is part of the practice of waste sorting.  The phrase "will be recycled" — Indicates the process of processing materials after their collection, which also relates to waste sorting.  The expression "for reuse" — Emphasizes the idea of resource reutilization, which is a key aspect of environmental practice.  Together, these elements create a clear understanding that the text refers to waste sorting and its subsequent recycling.</p>

Table 4: Examples of LLMs' output. The texts are given in English (translated from Russian).

First, one domain expert performed a detailed review of the model-generated explanations, providing structured comments on their adequacy, alignment with annotation guidelines, and typical reasoning patterns.

In the second stage, a second expert reviewed the initial assessment. In cases of uncertainty or disagreement, the experts discussed the instance jointly until a consensus was reached.

This procedure allowed us to ensure consistency of judgments while preserving detailed qualitative analysis. While the evaluation was not designed as a formal inter-annotator agreement study, the consensus-based review provides a reliable basis for identifying recurring strengths and limitations of

the explanations.

The goal of this section is to document how experts perceive model reasoning in relation to their own annotation principles.

## 5.1. Contextual Limitations of the Model

The main difficulty for LLMs is the lack of context, which directly affects its ability to reconstruct expert reasoning behind implicitly expressed practices (RQ1). If there is no information about the purpose of stores like "Mr. Thank You" ("Mr. Spasibo") and "Laugh-ka" ("Laffka"), it is very hard to understand why people bring their items there. While it is obvious to an expert that this is a practice of shared

use, the model starts to get confused, because it has to rely on the meaning of words in a short phrase, which does not always lead to correct conclusions. Another example is eco-friendly fashion. An expert knows how this practice addresses environmental problems and what established meanings are associated with it. On the other hand, the model can only guess what the word “eco-friendly” means in this phrase. When analyzing mistakes in such situations, LLMs point to the lack of context, insufficient information, and the formal nature of language analysis. The explanations of LLMs are especially accurate when it understands the context, for example, when it knows what is “plasticwatching” as an eco-practice.

## 5.2. Mismatch Between Expert Decisions and Model Reasoning

Another difficulty comes from expert decisions and highlights the mismatch between expert conceptual boundaries and model-internal associations (RQ3). For instance, experts decide that everything that happens before transporting separated waste to a recycler is waste sorting, which includes studying labels. The model, however, perceives the process as a whole: studying labels – sorting – transportation – recycling. It is difficult for it to distinguish where sorting ends and recycling begins, especially if there is no explicit statement about it. Another example of such an expert decision is the mention of volunteers. Experts assume that volunteers are involved in eco-practices, so their presence indicates the promotion of responsible consumption. At the same time, for the model volunteers are just people, so it needs additional information to make the conclusion. When analyzing expert judgment and model annotation, LLMs also do not have any information about the experts’ decisions, so they tend to agree with the model based on formal features.

For the analysis, LLMs refer to elements of the text in order to draw a conclusion about the validity of an expert’s or the model’s judgment. If it knows the context, it uses it.

## 5.3. Observed Weaknesses and Practical Value of Model Explanations

What can be concluded as drawbacks of the explanations of LLM:

1. The LLM approaches the analysis formally. For example, the word “free” is interpreted as refusing to buy something, because one cannot pay or purchase it;
2. LLMs may start analyzing elements of the text that are completely unrelated to the eco-

practice and are not significant from the expert’s point of view;

3. Complex actions, for example, the listing of the main activities of an environmental organization cannot be united by the LLM under a common meaning; it examines them separately;
4. LLMs do not always understand the task and begins to explain either the expert’s opinion or the model’s reasoning, pointing out what it considers weak, misleading, or, on the contrary, justified, or it proves the eco-friendliness of a practice rather than justifying its specific type;
5. LLMs will most likely make a mistake if it does not have access to all the information, for example, if it cannot see a photo of the packaging, it might conclude that the text is about studying labels, while there are no labels at the photo, only the specific type of the waste, such as polypropylene egg cartons;
6. LLMs may have overly high requirements for annotation, demanding exhaustive information about the practice, a full picture of it, which, of course, cannot be contained in a single phrase.

Why LLMs should be used for such tasks:

- From the point of view of the authors of the content, LLMs offer advice on how to convey a message more accurately, it helps to critically evaluate the wording in post texts, suggesting ways to make them more complete.
- From the expert’s point of view, LLMs help to identify the mistakes or inaccuracies in the annotation, see new interpretations of practices or to expand the range of practices, which are described in a given passage.
- From the programmer’s point of view, LLMs suggest how to fine-tune the classifier in order to avoid mistakes in the future, leading to the conclusion that training on the context is necessary.

## 6. Conclusion

The obtained results demonstrate that LLM explanations help reconstruct implicit expert reasoning (RQ1), diagnose systematic classification errors (RQ2), and reveal mismatches between expert annotation principles and model reasoning patterns (RQ3). Our findings are based on a diagnostic sample and should be interpreted as exploratory rather than statistically generalizable.

The analysis highlights that many errors stem from systematic semantic associations and overgeneralization rather than random noise. At the same time, LLM-generated explanations expose implicit assumptions embedded in both model predictions and expert annotations, thereby functioning as a diagnostic interface between computational models and domain knowledge.

More broadly, our study suggests that explanation-based analysis is a productive methodological tool for environmental NLP. By combining multi-label modeling with qualitative examination of generated rationales, researchers can better understand how sustainability discourse is represented, inferred, and sometimes misinterpreted by language models.

Future work may explore incorporating structured contextual knowledge into classification models, refining annotation guidelines based on explanation-driven feedback, and evaluating the faithfulness of generated explanations beyond their plausibility.

## 7. Acknowledgment

This study was supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the Carbon Measurement Test Area in Tyumen' Region (FEWZ-2024-0016).

## 8. Bibliographical References

### References

- Raven Adam and Marie Kogler. 2025. [Bidirectional topic matching: Quantifying thematic intersections between climate change and climate mitigation news corpora through topic modelling](#). In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 208–217, Vienna, Austria. Association for Computational Linguistics.
- Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. [Carbontracker: Tracking and predicting the carbon footprint of training deep learning models](#). ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems. ArXiv:2007.03051.
- Christina Barz, Melanie Siegel, and Daniel Hanss. 2025. [Analyzing the online communication of environmental movement organizations: NLP approaches to topics, sentiment, and emotions](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 68–76, Tallinn, Estonia. University of Tartu Library.
- Valentina Tretti Beckles and Adrian Vergara Heidke. 2025. [Thematic categorization on pineapple production in Costa Rica: An exploratory analysis through topic modeling](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 44–55, Tallinn, Estonia. University of Tartu Library.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent hatred: A benchmark for understanding implicit hate speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Carlo Galli, Anna V. Gavrilova, and Elena Calciolari. 2025. [Large language models in systematic review screening: Opportunities, challenges, and methodological considerations](#). *Information*, 16(5).
- Anna Glazkova and Olga Zakharova. 2025. [From data to grassroots initiatives: Leveraging transformer-based models for detecting green practices in social media](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 1–9, Tallinn, Estonia. University of Tartu Library.
- Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Ranran Haoran Zhang, Sujeeth Reddy Vummanthala, et al. 2024. [Evaluating LLMs at detecting errors in LLM responses](#). *arXiv preprint arXiv:2404.03602*.
- Yova Kementchedjhieva and Ilias Chalkidis. 2023. [An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text](#). In *Findings of the Association for Com-*

- putational Linguistics: ACL 2023*, pages 5828–5843.
- Arianna Longo and Alessandro Y. Longo. 2025. [No AI on a dead planet: Sentiment and emotion analysis across Reddit communities on AI and the environment](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 77–83, Tallinn, Estonia. University of Tartu Library.
- Natalia Loukachevitch, Natalia Tkachenko, Anna Lapanitsyna, Mikhail Tikhomirov, and Nicolay Rusnachenko. 2025. [RuOpinionNE-2024: Extraction of opinion tuples from russian news texts](#). In *Proceedings of the International Conference "Dialogue"*, volume 2025.
- Valentin Mamedov, Evgenii Kosarev, Gregory Leytner, Ilya Shchuckin, Valeriy Berezovskiy, Daniil Smirnov, Dmitry Kozlov, Sergei Averkiev, Lukyanenko Ivan, Aleksandr Proshunin, Ainur Israfilova, Ivan Baskov, Artem Chervyakov, Emil Shakirov, Mikhail Kolesov, Daria Khomich, Daria Latortseva, Sergei Porkhun, Yury Fedorov, Oleg Kutuzov, Polina Kudriavtseva, Sofii Soldatova, Kolodin Egor, Stanislav Pyatkin, Dzmitry Menshykh, Grafov Sergei Iurevich, Eldar Damirov, Vladimir Karlov, Ruslan Gaitukiev, Arkadiy Shatenov, Alena Fenogenova, Nikita Savushkin, and Fedor Minkin. 2025. [GigaChat family: Efficient Russian language modeling through mixture of experts architecture](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 93–106, Vienna, Austria. Association for Computational Linguistics.
- Sarah Masud, Ashutosh Bajpai, and Tanmoy Chakraborty. 2025. [Focal inferential infusion coupled with tractable density discrimination for implicit hate detection](#). *Natural Language Processing*, 31(6):1323–1349.
- Telma Peura, Attila Krizsán, Salla-Riikka Kuusalu, and Veronika Laippala. 2025. [Perspectives on forests and forestry in Finnish online discussions - a topic modeling approach to suomi24](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 10–15, Tallinn, Estonia. University of Tartu Library.
- Saptarshi Sengupta, Wenpeng Yin, Preslav Nakov, Shreya Ghosh, and Suhang Wang. 2025. [Exploring language model generalization in low-resource extractive QA](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7106–7126, Abu Dhabi, UAE. Association for Computational Linguistics.
- Dmitrii Stoianov, Danil Taranets, Olga Tsymboi, Ramil Latypov, Almaz Dautov, Vladislav Kruglikov, Nikita Surkov, German Abramov, Pavel Gein, Dmitry Abulkhanov, et al. 2025. [T-pro 2.0: An efficient Russian hybrid-reasoning model and playground](#). *arXiv preprint arXiv:2512.10430*.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. [Multilingual translation from denoising pre-training](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.
- Olga Zakharova and Anna Glazkova. 2024. [GreenRu: A Russian dataset for detecting mentions of green practices in social media posts](#). *Applied Sciences*, 14(11):4466.
- Olga V Zakharova, Anna V Glazkova, Irina N Pupyshova, and Natalia V Kuznetsova. 2022. [The importance of green practices to reduce consumption](#). *Changing Societies & Personalities*. 2022. Vol. 6. Iss. 4, pages 884–905.
- Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, Jacob Devasier, and Chengkai Li. 2024. [Granular analysis of social media users' truthfulness stances toward climate change factual claims](#). In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 233–240, Bangkok, Thailand. Association for Computational Linguistics.
- Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. [Explainability for large language models: A survey](#). *ACM Trans. Intell. Syst. Technol.*, 15(2).
- Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. [Can large language models understand context?](#) In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2004–2018, St. Julian's, Malta. Association for Computational Linguistics.