LLMs to the Rescue: Explaining DSA Statements of Reason with Platform's Terms of Services

Marco Aspromonte[1]
marco.aspromonte2@unibo.it

Andrea Filippo Ferraris[1]
andrea.ferraris3@unibo.it

Federico Galli[1,2]
federico.galli7@unibo.it

Giuseppe Contissa[1,2]
giuseppe.contissa@unibo.it

[1] **Alma AI, Alma Mater Studiorum, University of Bologna**
[2] Department of Legal Studies, Alma Mater Studiorum, University of Bologna

Abstract

The Digital Services Act (DSA) requires online platforms in the EU to provide "statements of reason" (SoRs) when restricting user content, but their effectiveness in ensuring transparency is still debated due to vague and complex terms of service (ToS). This paper explores the use of NLP techniques, specifically multi-agent systems based on large language models (LLMs), to clarify SoRs by linking them to relevant ToS sections. Analysing SoRs from platforms like Booking.com, Reddit, and LinkedIn, our findings show that LLMs can enhance the interpretability of content moderation decisions, improving user understanding and engagement with DSA requirements.

1 Introduction

The Digital Services Act (DSA), adopted by the European Union on November 1, 2022, represents a significant milestone in the EU regulation of online platforms, as it establishes a global standard for transparency and accountability in content moderation.

A key innovation of the DSA is the requirement for intermediary hosting services to provide "statements of reason" (SoRs) when restricting user-generated content (Article 17). The SoR must specify the action taken, the factual circumstances, any use of automated systems in the moderation process, and the legal or contractual grounds for deeming the content illegal or incompatible with the platform's terms of service (ToS), along with an explanation and other metadata. It should also inform users of available redress options, ensuring clarity and precision to allow users to contest the decision.

Article 24(5) further requires online platforms to submit SoRs, as outlined in Article 17, to the European Commission for inclusion in a publicly accessible, machine-readable database. In response to this mandate, the European Commission launched the DSA Transparency Database (TD) in September 2023.

The scheme of the TD roughly reflects the content of the SoR pursuant to Article 17.[1] Each SoR instance is composed of several mandatory attributes, such as the content type (e.g. text, image, etc.) and language, the type and period of restriction, the ground for the decision, the category of restricted content, the fact relied upon on the decision, etc. Attribute values are to be selected by the provider from a list of options or can be typed into as free text (generally with character limitations). Other attributes are only optional. As for the ground for the provider's decision (field "decision_ground"), the TD presents only two possible options: `"ILLEGAL_CONTENT"` and `"INCOMPATIBLE_CONTENT"`. Moreover, the TD typifies 14 distinct "categories" of statements as potential grounds for restriction (see Table 1).

The TD is intended as a critical tool for scrutinizing content moderation practices, revealing how well platforms comply with the requirements set by the DSA. However, the effectiveness of the TD in fulfilling its promises of transparency and accountability remains a subject of ongoing debate (Trujillo et al., 2024; Kaushal et al., 2024). In particular, there are doubts about whether the SoR provides sufficient information to allow users to understand the reasons for content restriction and to contest its lawfulness. This is especially true when it comes to the asserted incompatibility with ToS. These are often lengthy and complex documents

[1] Available at: transparency.dsa.ec.

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 205–215
November 16, 2024 ©2024 Association for Computational Linguistics

(Melinat et al., 2014) drafted in legal jargon (Butt, 2001), and the complexity and lack of understanding and awareness of legal texts of this kind is an old and well-known issue (Masson and Waldron, 1994).

Given this background, the present paper explores the potential of NLP techniques, particularly multi-agent systems based on large language models (LLMs), to enhance the transparency and user-friendliness of SoRs submitted to the TD. We propose and evaluate a method that uses LLMs to contextualise explanations in SoRs related to ToS incompatibility within the platform's content policy guidelines. Our focus on ToS incompatibility stems from its frequency as the main reason for content removal and its suitability for uniform LLM-based analysis, unlike removals based on national laws, which vary across the EU and add complexity. ToS, being unique to each platform, offer a more consistent and manageable basis for explanation.

2 Background

Our work builds on and merges three emerging research strands: 1) existing works applying computational techniques to analyse the DSA Transparency Report; 2) the application of NLP techniques to enhance accessibility and legibility of transparency legal requirements; 3) the use of LLMs in the legal field.

2.1 Computational Analysis of DSA Transparency Database

Since the launch of the DSA Transparency Database, several studies have used computational methods to analyze and extract aggregated insights from its data. The database contains vast amounts of raw data on content moderation practices by online platforms, making automated tools essential for understanding its contents effectively.

For instance, Drolsbach et al. (Drolsbach and Pröllochs, 2024) examined 156 million SoRs over two months, highlighting content restrictions categorized under "Scope of Platform Service" (49.06%), reflecting ambiguities in this classification. Similarly, another work (Trujillo et al., 2024) analysed 195 million SoRs, incorporating cross-references with Article 15 Transparency Reports[2]. They found inconsistencies across platforms like TikTok, YouTube, and Snapchat.

A key related study (Kaushal et al., 2024) to our paper analysed a representative sample of the Transparency Database (131m SoRs) submitted in November 2023 to evaluate platform content moderation practices. They provided several findings, such as the prevalence of SoR reported as ToS violations (99.8%) compared to illegal content (0.2%). They show that all (99.9%) of ToS violations do not report the URL to the relevant platforms' ToS. With regard to ToS, they also point to a critical lack of precision in stating the "fact underlying the decision", namely the motivation of the decision taken. This does not generally allow users to identify what elements of their content are violating norms, leading to restriction.

Overall, current research indicates that online platforms heavily rely on their ToS as the basis for content restriction decisions, which is, per se, compliant with the DSA. However, when content is deemed incompatible with the ToS, the communication often lacks specificity, providing only a generic statement without a clear reference to the exact grounds for removal. In our study, we investigate whether NLP techniques can be employed to link ToS to the relevant sections of online platforms' ToS or content guidelines, thereby giving users more detailed information about the reasons for content restrictions.

2.2 NLP for Legal Transparency Enhancement

NLP offers significant potential for enhancing transparency and regulatory compliance in the legal domain (Thimm, 2023; Cejas et al., 2023). By automating the analysis and generation of complex legal texts, NLP can improve business compliance, reduce human errors and improve the clarity of legal communications (Katz et al., 2023). This capability is particularly valuable in contexts where legal requirements are intricate and frequently updated - such as the digital environment - ensuring that organisations can maintain compliance consistently and transparently (Zhou et al., 2022).

For example, NLP can be employed to automatically extract relevant clauses from regulatory documents and cross-reference them with a company's internal policies to ensure alignment with legal standards (Bizzaro et al., 2024; Hendrycks et al., 2021). In another scenario, NLP tools can analyse public statements or contractual terms to identify potential legal risks and unfair clauses and enable proactive compliance management (Lippi et al., 2019).

[2] Article 15 of the DSA mandates annual transparency reports from platforms on content moderation actions and their justifications.

Furthermore, NLP can play a crucial role by making complex legal language more accessible to users (Garimella et al., 2022). Automated systems can translate intricate legal jargon into plain language, helping users understand the rationale behind moderation decisions and, if necessary, challenge those decisions effectively. This not only enhances user engagement but also builds trust in platform governance by providing transparency into the legal reasoning that underpins content moderation.

2.3 LLMs for Legal Applications

LLMs are rapidly transforming legal practice by automating complex tasks such as interpreting legal texts, generating documents, and providing preliminary legal advice (Qin and Sun, 2024; Yang et al., 2024; Martin et al., 2024). These models are particularly valuable in domains that involve the processing of large volumes of intricate and nuanced language, offering the potential to significantly enhance both transparency and understandability, and efficiency in various legal processes.

The application of LLMs is notably expanding across various legal domains. They are increasingly employed to draft legal documents that comply with specific regulatory requirements (Lin and Cheng, 2024), automate the extraction of relevant clauses from extensive legal texts (Bizzaro et al., 2024), and even predict the outcomes of legal disputes based on historical data (de Menezes-Neto and Clementino, 2022). This growing interest highlights the transformative role LLMs can play in streamlining legal processes, which are traditionally reliant on significant human expertise and time.

In the context of content moderation, LLMs show considerable promise as tools both for supporting platforms in their content moderation activity (Kumar et al., 2024; Kolla et al., 2024) as well as for helping users understand and, if necessary, challenge platform decisions (Guan et al., 2023). By analysing Statements of Reason provided when content is removed or restricted, LLMs can leverage their advanced NLP capabilities to interpret SoRs and assess whether moderation actions comply with the DSA and platform-specific ToS (Atreja et al., 2023).

3 Data

For this study, we compiled a custom dataset using resources from the DSA Transparency Database.

We focused on Statements of Reasons (SoRs) specifically related to content removal due to violations of terms of services (ToS) from three major online platform providers: Booking.com, Reddit, and LinkedIn. This selection was made to capture a diverse range of online environments. Booking.com, as a leading e-commerce platform in the travel industry, provides insights into ToS enforcement concerning commercial content; Reddit, a large social media forum, illustrates content moderation challenges in a user-generated, community-driven space; and LinkedIn, a professional networking platform, reflects ToS enforcement in a setting focused on professional conduct and business communication. This approach enables a comprehensive examination of SoRs across platforms with varying purposes, user bases, and content policies.

To ensure a representative sample, we selected SoRs from a specific time frame, spanning from March 2024 to August 2024.

The content of each SoR consists of four key attributes included in the TD, which are intended to provide context for explaining the decision that affects users' content. The attributes are the UUID, the ground for incompatible content (`"incompatible_content_ground"`), the explanation for incompatible content (`"incompatible_content_explanation"`) and the facts relevant for the decision (`"decision_facts"`).

We did not rely on the current versions of the ToS available on the platforms' websites, as they may overlook regulatory changes or evolving industry standards that could impact the interpretation of the SoRs. Using the historical ToS in force at the time the SoRs were issued was crucial, as relying on updated versions could render certain SoRs outdated or irrelevant.

The selected dataset consists of 7000 SoRs, among which 3000 were issued by Booking.com, 2000 by LinkedIn and 2000 by Reddit.

Figure 1 shows the distribution of the selected SoRs across the three online platforms and the 14 typified restrictions, while Table 1 represents the pairs between each category and its acronym. LinkedIn and Booking.com have high scores in the "Scope of Platform Service" (SOPS) category, reflecting their specific and well-defined content purposes. Booking.com also shows a high number of SoRs for "Data Protection and Privacy Violations" (DPAPV) due to its frequent handling of sensitive user data.

Reddit scores highest in "Non-Consensual Be-

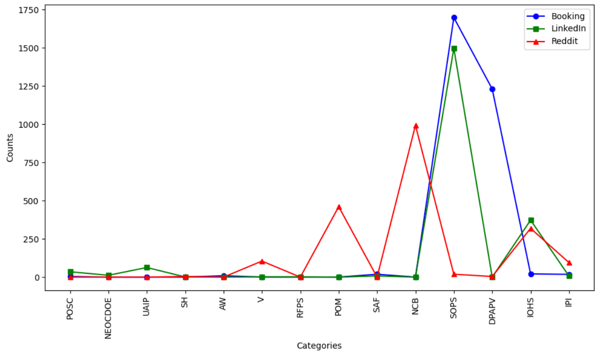

Figure 1: Distribution of categories in linear scale

Category	Acronym
PORNOGRAPHY_OR_SEXUALIZED_CONTENT	POSC
NEGATIVE_EFFECTS_ON_CIVIC_DISCOURSE_OR_ELECTIONS	NEOCDOE
UNSAFE_AND_ILLEGAL_PRODUCTS	UAIP
SELF_HARM	SH
ANIMAL_WELFARE	AW
VIOLENCE	V
RISK_FOR_PUBLIC_SECURITY	RFPS
PROTECTION_OF_MINORS	POM
SCAMS_AND_FRAUD	SAF
NON_CONSENSUAL_BEHAVIOUR	NCB
SCOPE_OF_PLATFORM_SERVICE	SOPS
DATA_PROTECTION_AND_PRIVACY_VIOLATIONS	DPAPV
ILLEGAL_OR_HARMFUL_SPEECH	IOHS
INTELLECTUAL_PROPERTY_INFRINGEMENTS	IPI

Table 1: Categories and Their Acronyms

havior" (NCB), likely due to its large, diverse user base and the anonymity it offers, which can lead to issues like doxxing, harassment, and the unauthorised sharing of personal information. On the contrary, categories such as Animal Welfare (AW), Self-Harm (SH), and Risk for Public Security (RFPS) have relatively low SoR frequencies across all platforms, as these are less common restrictions in the contexts analysed. Overall, Reddit addresses the most diverse harmful content, while LinkedIn and Booking.com focus on specific issues related to their platform's nature.

In addition to the Statements of Reasons (SoRs), we collected the relevant Terms of Service (ToS) in effect when the selected SoRs were issued to understand the basis for content removals. We relied on both the ToS, as the binding contract, and the community guidelines, which provide additional context for applying the ToS. Though not part of the formal contract, community guidelines are valuable as they offer practical interpretations of the ToS. Integrating both allowed us to better align the explanations of content removals with their intended meaning and scope, providing context that might be missing from the ToS alone.

Both the ToS and community guidelines were pre-processed to extract relevant content moderation clauses, enabling the LLM to match them with the SoRs for more accurate and contextually relevant explanations of the moderation decisions.

4 Architecture and Methods

The proposed architecture is based on a multi LLM-based-agent system (Guo et al., 2024) and a Retrieval-Augmented Generation (RAG) process (Gao et al., 2024). It employs two autonomous LLM-based agents, each assigned specific roles: the "Refiner Agent" and the "Explainer Agent". These agents operate independently, coordinating their actions to process and interpret platform documents (ToS and SoRs), enhancing both the accuracy and contextual relevance of the system's output. Through this division of tasks and inter-agent interaction, our approach aligns with the principles of multi-agent systems by enabling collaborative decision-making and specialised behaviour.

We tested the agent-based architecture with two pre-trained LLMs: `"Mistral-7b-instruct-v0.3"` (Jiang et al., 2023), and `"Gpt4o-mini"` [3]. During each test run, only one of these models is used, enabling a direct comparison of their outputs. Each model is independently evaluated for its ability to interpret retrieved documents, refine them, and generate expert-like explanations.

The Mistral-7b model was used in an optimised version with 4-bit quantisation, which allows it to handle complex prompts efficiently while minimising memory usage. This makes it suitable for resource-constrained environments (Pan et al., 2023). On the other hand, the GPT4o-mini model has a unique architecture that adds additional depth and nuance to the evaluation process.

These models were integrated into the architecture using the Hugging Face Transformers library [4] for the Mistral model and the OpenAI API for the GPT4o-mini model. This integration enables comprehensive performance assessments across different computational scenarios. By combining the transparency and replicability of open-source models with the enhanced performance of proprietary models, this dual approach facilitates a thorough comparison of the models' effectiveness in interpreting and evaluating content moderation actions.

The architecture is hosted on a public GitHub

[3] See: Gpt-4o-mini-OpenAI

[4] See: huggingface.co

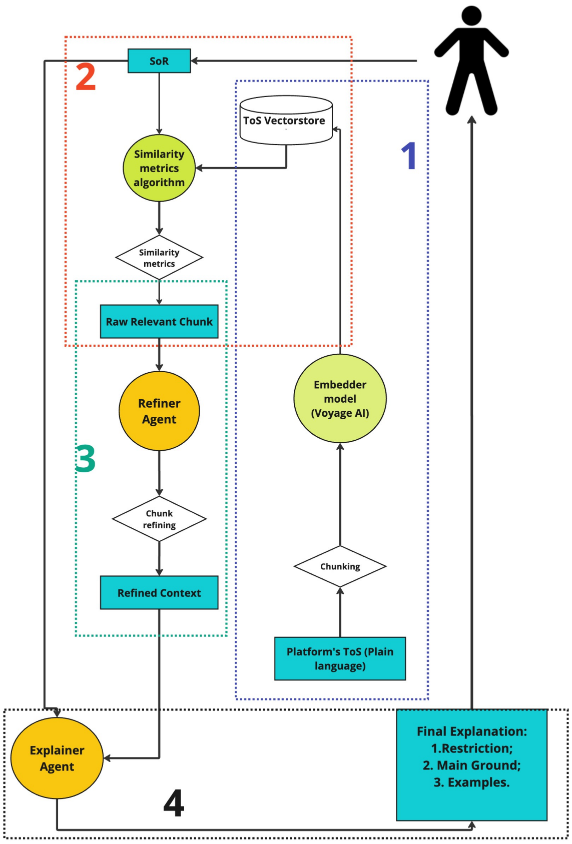

Figure 2: Visual representation of the pipeline

repository [5] and is presented in Figure 2, and it is divided into 4 modules, each marked with a coloured box and pairing number.

- Module 1 (blue): Vector Store Creation;
- Module 2 (red): Retriever and similarity;
- Module 3 (green): Agentic refinement;
- Module 4 (black): Agentic explanation.

We analyse each module in the following subsections.

4.1 Vector Store Creation

The first module, computed once for each online platform provider, is designed to create a chunked version of the relevant ToS. We achieved this by dividing the ToS into chunks, ensuring each chunk corresponds to a complete paragraph or section, thereby preserving the text's original structure and semantic meaning.

We then initialised an embedding model using `VoyageAIEmbeddings`[6], which converts the text into high-dimensional vectors within a dense vector space to effectively capture its semantic meaning. Specifically, we utilised the `Voyage-2-Law` large pre-trained embedding model,[7] which is tailored for legal texts. We opted for Voyage-2-Law over general-purpose models because it is specifically trained on legal documents, enabling it to capture the nuances and context of legal language more accurately.

The generated embeddings were stored in the open-source vector database, Chroma DB (Chroma).[8] These chunked ToS serve as a knowledge base, facilitating the retrieval of relevant sections of the ToS in relation to a given SoR.

4.2 Retriever and Similarity

The primary goal of this module is to extract from the database all ToS chunks that are relevant to the given SoR. It begins by analysing the SoR and focuses on retrieving the most semantically relevant chunks from the vector store. To achieve this, we adopted a hybrid approach[9], combining Cosine Similarity metrics (a semantic-based method) (Lahitani et al., 2016) with the Probabilistic Relevance framework (specifically, BM25) (Robertson and Zaragoza, 2009).

We selected the top two results from each method and merged them into a single file consisting of a list of chunks (referred to as "Raw Relevant Chunks" in 2). To avoid redundancy, we opted to filter out the identical chunks from the file "Raw Relevant Chunks" in case the chunks retrieved using Cosine Similarity overlap with those obtained via BM25. This process results in a list that may contain only two chunks. The file is then passed to the next module, the Agentic Refiner, for further processing.

4.3 Agentic Refinement

The third module focuses on refining the chunks extracted by the second module to streamline and optimise the information that will be provided to the agent responsible for generating the user explanation. Following a novel approach explored by Xu et al. (2024), we employ the first LLM-based agent to refine the chunks obtained from the previous module.

[5] See: framework's GitHub repository

[6] See: `https://www.voyageai.com/`

[7] See: Voyage-2-Law overview

[8] See: `https://github.com/chroma-core/chroma`

[9] See: Hybrid Search: Combining BM25 and Semantic Search

In this module, the agent is provided with the target SoR and the "Raw relevant chunks". Using techniques of prompt engineering (Sahoo et al., 2024), the agent is instructed to extract only the information from the raw chunks that directly relates to the target SoR, removing any irrelevant content and eliminating noise that may be present in the raw data.[10]

4.4 Agentic Explanation

The fourth and final module, the Agentic Explanation module, is responsible for explaining the SoR in relation to the platform's ToS. Drawing on the work of Feng et al. (2023), which highlights the effectiveness of large language models (LLMs) in rephrasing and simplifying complex legal texts, this module utilizes the SoR and the refined sections of the ToS to link the moderation action to the platform's contractual justifications.

The output provides a structured explanation to enhance users' understanding of the legal grounds for content moderation. The agent situates the SoR within the platform's policy framework by identifying the ground or rule that the content violated (rule-based explanation) and offering examples to demonstrate how the ground applies to different forms of content (explanation by example) (van der Waa et al., 2021).

It is important to note that this kind of explanation does not extend to the platform's internal decision-making process or the criteria used to assess a particular content for restriction. This limitation is due to the fact that the TD does not provide data on the actual moderated content. As a result, the agent cannot explain why a particular piece of content was deemed problematic under the platform's rules or account for any contextual factors influencing the moderation decision. Nonetheless, the output can still help users who are already familiar with the content in question to better understand the reasons behind the restriction.

5 Validation

The validation process focuses on evaluating the performance of two LLM-based agents: the "Refiner", which extracts relevant sections from the ToS, and the "Explainer", which aims to clarify the content of a SoR in light of the relevant platform's ToS.

A human evaluation approach was chosen to assess the quality of the outputs generated by both agents. Human evaluation was selected due to its capacity to provide a nuanced and contextual assessment that goes beyond what current automated metrics can offer (Chang et al., 2024). It allows for more accurate and comprehensive feedback on semantic and qualitative aspects of the generated responses, which is particularly important with legal content.

The evaluation process was designed to achieve statistical significance, ensuring that the results are robust and credible. In particular, we observed that, within each provider, the SoRs pertaining to the same category are remarkably similar in their formulation. Given this high degree of standardisation or consistency per category, we selected one representative sample from each category per online platform.

The criteria for validating the outputs of the two agents were based on four key metrics, each rated on a 1-to-5 scale:

1. **Relevance**: Assesses whether the output is appropriate and significant with regard to the Statement of Reasons (SoR) and the refined ToS. High scores indicate that the refined content is directly relevant to the SoR/refined ToS, while lower scores suggest a lack of alignment or relevance.

2. **Accuracy**: Evaluates whether all relevant arguments and information from the original ToS (for the "Refiner") and the refined ToS (for the "Explainer") are retained. A high score reflects comprehensive retention, whereas a low score indicates omissions.

3. **Coherence**: Measures the consistency of the output with the original ToS intent. Under this metric, the "Refiner" is evaluated in terms of linguistic coherence, namely its capacity to faithfully represent original text. The "Explainer" is assessed in terms of its capacity not to hallucinate and introduce meanings and examples which are not directly taken by the refined ToS. Higher scores signify that the output faithfully reflects the original content without modification/hallucination.

4. **Readability** (specific to the "explainer"

[10] The prompt includes specific instructions to the model, detailing the background and context for evaluation. It ensures that the model considers the statement of reason provided by the platform, the relevant sections of the ToS, and the legal framework context outlined in Article 17.

agent): Assesses the clarity and ease of understanding of the generated explanations. Higher scores suggest that the output is easy to read, with a smooth flow and consistent tone and style.

The evaluation was conducted by a panel of three independent human evaluators, each with specialised expertise in content moderation practices and regulatory compliance under the DSA[11].

6 Experimental Results

This section presents the experimental results obtained from evaluating the two LLM-based agents, the "Refiner Agent" and the "Explainer Agent", using the selected pre-trained models: "Mistral-7b-instruct-v0.3" and "GPT4o-mini".

Table 2 summarises the performance metrics for the Refiner Agent across the different criteria and analysed platforms.

The Explainer Agent was evaluated separately to measure its effectiveness in providing user-friendly explanations that contextualise the legal reasons behind content moderation decisions. Table 3 shows the performance metrics for the Explainer Agent.

The scores (1-5) were averaged for both agents, outlining a global statistical significance and providing a clear comparison of the models' outputs.

Platform	**Model**	**Relevance**	**Accuracy**	**Coherence**
Booking.com	GPT4o-mini	**4.69**	3.84	4.38
Booking.com	Mistral-7b	4.07	**4.28**	4.5
Reddit	GPT4o-mini	4.45	3.80	4.60
Reddit	Mistral-7b	4.0	4.05	4.5
LinkedIn	GPT4o-mini	4.56	4.0	**4.68**
LinkedIn	Mistral-7b	3.81	3.75	4.37

Table 2: Results for Refiner Agent across platforms

Platform	**Model**	**Relevance**	**Accuracy**	**Coherence**	**Readability**
Booking.com	GPT4o-mini	**4.85**	4.57	**4.85**	4.71
Booking.com	Mistral-7b	4.71	**4.73**	4.14	4.9
Reddit	GPT4o-mini	4.71	4.12	4.62	**5.0**
Reddit	Mistral-7b	4.7	3.8	4.2	4.73
LinkedIn	GPT4o-mini	4.0	4.4	4.7	4.8
LinkedIn	Mistral-7b	4.75	4.0	4.12	4.62

Table 3: Results for Explainer Agent across Platforms

We used standard deviation to quantify the variability in the scores provided by different evaluators across the relevant metrics. To facilitate comparison across different metrics and model/platforms, we also normalised the standard deviation values to a range a range $[0, 1]$.

Table 4 and Table 5 present the standard deviation values across the different criteria per model-platform, respectively for the Refiner and the Explainer Agent. The lower variability scores show the higher inter-annotator agreement.

Platform	**Model**	**Relevance**	**Accuracy**	**Coherence**
Booking	gpt4mini	**0.29**	0.43	0.30
Booking	mistral-7b	0.58	**0.41**	0.46
Reddit	gpt4mini	0.31	0.49	**0.23**
Reddit	mistral-7b	0.43	0.47	0.25
LinkedIn	gpt4mini	0.39	0.47	0.34
LinkedIn	mistral-7b	0.46	0.47	0.43

Table 4: Standard Deviation for Refiner Agent

Platform	**Model**	**Relevance**	**Accuracy**	**Coherence**	**Readability**
Booking	gpt4mini	0.30	0.38	0.41	0.30
Booking	mistral-7b	0.30	0.44	0.44	0.51
Reddit	gpt4mini	**0.00**	0.36	**0.22**	**0.27**
Reddit	mistral-7b	0.25	0.40	0.20	0.39
LinkedIn	gpt4mini	0.40	**0.28**	0.23	**0.27**
LinkedIn	mistral-7b	0.43	0.49	0.42	0.47

Table 5: Standard Deviation for Explainer Agent

7 Discussion

We detail the discussion in the subsections below, separately for the two agents and then comparatively on the performance of the two models.

7.1 Refiner Agent Results

The Refiner Agent was evaluated on relevance, accuracy, and coherence. The results across platforms show notable differences between the two LLM models used — `GPT4o-mini` and `Mistral-7b`.

For relevance, `GPT4o-mini` generally outperformed Mistral-7b across all platforms, achieving the highest scores on Booking.com (4.69) and LinkedIn (4.56), with strong evaluator agreement indicated by low standard deviations (0.29 on Booking.com and 0.23 on Reddit). This indicates gpt4o-mini's ability to retrieve the most relevant sections of the ToS for the given Statement of Reasons (SoR). Mistral-7b, though slightly lower in relevance scores, still performed consistently, particularly on Booking.com (4.07) and Reddit (4.0).

Accuracy scores demonstrate that Mistral-7b surpassed GPT4o-mini in most cases, particularly on Booking.com (4.28) and Reddit (4.05). This suggests that Mistral-7b performed better at retaining and faithfully representing the necessary arguments from the original ToS. However, the higher standard deviations for Mistral-7b in coherence (0.46 on Booking.com and 0.43 on LinkedIn) suggest

[11] The dataset and the evaluation results can be found at the following GitHub repository: `https://github.com/sustaz/DAFNE_4_NLLP`

more inconsistent outputs in terms of logical structure and clarity. Also, on LinkedIn, GPT4o-mini performed better (4.0), possibly due to the platform's more structured and formal ToS, which may have aligned better with its training data.

In terms of coherence, GPT4o-mini again showed stronger results, particularly on LinkedIn (4.68) and Reddit (4.6), suggesting its capacity to maintain a logical flow in refining the ToS. Mistral-7b was slightly lower but still consistent, scoring 4.5 on both Booking.com and Reddit.

7.2 Explainer Agent Results

The Explainer Agent was evaluated on four metrics: relevance, accuracy, coherence, and readability. Similar trends emerged across the platforms, with GPT4o-mini showing the strongest performance in most categories, particularly in readability.

For relevance, GPT4o-mini achieved the highest scores, especially on Booking.com (4.85) and Reddit (4.71), with perfect evaluator agreement on Reddit (0.00). Mistral-7b performed comparably well on Booking.com (4.71) and Reddit (4.7) and even surpassed GPT4o-mini on LinkedIn (4.75), but with the highest variability in terms of inter-annotator agreement.

Accuracy scores followed a similar trend, with Mistral-7b outperforming GPT4o-mini on Booking.com (4.73), but showing greater variability in readability with higher standard deviations on LinkedIn (0.47) and Booking.com (0.51). However, GPT4o-mini performed better on Reddit (4.12) and LinkedIn (4.4), with the lowest deviations between annotators (0.27), again suggesting its effectiveness in explaining less complex ToS.

All in all, the Explainer Agent, both with GPT4o-mini and Mistral-7b, scores higher in accuracy than the Refiner Agent. This suggests that, although it is useful for reducing context to explain the SoR with ToS, the refinement exercise may lead to the loss of information needed to provide a faithful explanation.

Coherence, i.e. whether the explanation logically connects the SoR to the ToS, showed GPT4o-mini excelling, particularly on Booking.com (4.85) and Reddit (4.62). This indicates the model's capability to provide clear, structured reasoning without introducing extraneous information.

In terms of readability, GPT4o-mini scored particularly high, achieving a perfect score on Reddit (5.0) and 4.71 on Booking.com. This suggests that GPT4o-mini generates explanations that are easier to read and understand, which is crucial for users engaging with the platform. Mistral-7b also performed well in readability, particularly on Booking.com (4.9), but showed a drop on Reddit (4.73), indicating slightly less user-friendly outputs.

The Box 7.2 portrays an example of an explanation of a Reddit's SoR produced with Mistral-7B.

Explanation Example

RESTRICTION: Your content was restricted because it violated Rule 3 of the platform's Terms and Conditions.
MAIN GROUND: The main ground for the restriction is the sharing of intimate or sexually-explicit media of someone without their consent.
EXAMPLE: Examples of content that the Platform may remove under this ground include:
- Images or videos of intimate parts of a person's body, even if the person is clothed or in public, if contextualized in a salacious manner (such as "creepshots" or “upskirt” imagery).
- Fake or “lookalike” depictions of intimate media.
- Leaked, stolen, or privately-shared content of someone who does not consent to sharing it on Reddit.
- Images or video of another person posted for the specific purpose of faking explicit content or soliciting “lookalike” pornography (e.g. “deepfakes” or "bubble porn").

7.3 Models Comparative Evaluation

Across the two agentic tasks, gpt4o-mini generally outperforms Mistral-7b across most metrics, particularly in relevance, coherence, and readability. This performance is further supported by lower standard deviations in these metrics, indicating greater consistency in GPT4o-mini's outputs. This can likely be attributed to gpt4o-mini's advanced architecture, which better handles the structured legal language in the ToS, making it well-suited for user-facing applications where clarity and accessibility are essential.

Mistral-7b, on the other hand, excels in accuracy, faithfully retaining details from the original ToS. This makes Mistral-7b a promising tool for tasks like legal document processing or back-end content moderation, where accuracy is key. However, Mistral-7b showed higher standard deviations in readability (e.g., 0.51 on Booking.com and 0.47 on LinkedIn), suggesting more variability in its user-friendliness.

Platform-specific variations further underscore the importance of ToS. For example, gpt4o-mini performed better on LinkedIn due to the structured nature of its ToS, while Mistral-7b excelled on Booking.com, where detailed ToS favoured accuracy.

Overall, both models produced useful outputs. Gpt4o-mini delivered more coherent and user-friendly explanations, which is ideal for front-end roles, while Mistral-7b prioritised accuracy, making it reliable for back-end tasks.

8 Limitations

We acknowledge a few limitations in our study, many of which stem from the inherent limitations of the DSA Transparency Database.

One major constraint is the absence of direct links or detailed descriptions of the moderated content. This limitation affects our system's ability to provide fact-specific explanations for content removal decisions. Instead, the model is forced to generate more generic, rule-based, or example-based explanations, which, despite their usefulness, can limit users' ability to fully understand how their content violated the platform's TOS.

Another limitation is the lack of multilingual testing. The models have only been tested on English-language data, as the database contains no non-English SoRs. Multilingual support is essential for broader applicability, especially across the EU.

Lastly, the system has not been evaluated on other platforms, which may provide different contexts for content restrictions and reasons for ToS violations. Such differences may impact the models' performance, and future work should address this by testing on more varied scenarios.

9 Conclusion and Future Work

This study demonstrated the potential of large language models, like GPT4o-mini and Mistral-7b, in enhancing transparency and user comprehension in content moderation decisions under the DSA.

However, challenges remain, particularly in handling the complexity of legal texts. LLMs struggle with nuanced, context-specific legal language (Homoki and Ződi, 2024), as the one used in ToS, as well as accuracy issues regarding reliance on static datasets, which may become outdated (Jayakumar et al., 2023).

Moreover, the "black box" nature of LLMs, where the decision-making process is opaque, poses a significant challenge in legal contexts (Lin et al., 2024). In legal applications, where the rationale behind decisions must be clear and defensible (Rotolo and Sartor, 2023), the inability to trace or explain the reasoning of LLMs undermines their reliability.

A valuable direction for future work is conducting an ablation study to better understand the contributions of various components in our system, particularly the role of the Refiner Agent. Preliminary results indicate that the Refiner Agent performs with slightly lower accuracy than the final model, prompting a closer examination of its role.

From a legal point of view, we intend to expand our work by linking the agentic explanations to more refined legal grounds for content removal contained in the ToS, possibly attaching them to relevant regulatory frameworks. The potential is not merely to provide an explanation of content restriction but also the legal justification to challenge the platform's decision.

10 Ethical Statement

There are several ethical strengths to our work. Data contained in the TD are anonymised. So, no personal data processing is involved in the study. The focus on explainability and transparency aims to empower platforms' users to better understand content moderation decisions in the context of ToS, possibly supporting their right to challenge the decision, contest the legality of ToS and seek redress.

Ethical concerns are related to the system's accuracy. Limited detail in platform reports can result in vague explanations. In some cases, the AI may "hallucinate" by generating incorrect or invented information which does not reflect the ToS content. These issues could mislead users and negatively impact their ability to effectively appeal decisions, potentially undermining their right to remedy.

Acknowledgement

This research was supported by the PRIN 2022 PNRR Project DAFNE (Democratic governance of Automated systems for Fake NEws, GA. P2022R7RS9) under the MUR National Recovery and Resilience Plan funded by the European Union – Next Generation EU.

References

Shubham Atreja, Jane Im, Paul Resnick, and Libby Hemphill. 2023. Appealmod: Shifting effort from moderators to users making appeals. *arXiv preprint arXiv:2301.07163*.

Pietro Giovanni Bizzaro, Elena Della Valentina, Maurizio Napolitano, Nadia Mana, and Massimo Zancanaro. 2024. Annotation and classification of relevant clauses in terms-and-conditions contracts. *Preprint*, arXiv:2402.14457.

Peter Butt. 2001. Legalese versus plain language. *Amicus Curiae*, 35:28.

Orlando Amaral Cejas, Muhammad Ilyas Azeem, Sallam Abualhaija, and Lionel C Briand. 2023. NLP-based Automated Compliance Checking of Data Processing Agreements against GDPR. *arXiv preprint arXiv:2209.09722v2*. Submitted to IEEE for possible publication.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Jacob de Menezes-Neto and M. B. M. Clementino. 2022. Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts. *PloS one*, 17(7):e0272287.

Chiara Patricia Drolsbach and Nicolas Pröllochs. 2024. Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 939–942.

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. *Preprint*, arXiv:2302.11957.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nanda Kambhatla. 2022. Text simplification for legal domain: Insights and challenges. In *Proceedings of the 1st Workshop on Natural Language Legal Processing (NLLP)*, pages 224–234.

Yanchu Guan, Dong Wang, Zhixuan Chu, Shiyu Wang, Feiyue Ni, Ruihua Song, Longfei Li, Jinjie Gu, and Chenyi Zhuang. 2023. Intelligent Virtual Assistants with LLM-based Process Automation. *Preprint*, arXiv:2312.06677.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *Preprint*, arXiv:2103.06268.

Péter Homoki and Zsolt Ződi. 2024. Large language models and their possible uses in law. *Research Article*, pages 435–455. Online Publication Date: 15 Apr 2024.

Thanmay Jayakumar, Fauzan Farooqui, and Luqman Farooqui. 2023. Large Language Models are Legal but they are not: Making the Case for a Powerful LegalLLM. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 223–229, December 7, 2023. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J. Bommarito II au2. 2023. Natural language processing in the legal domain. *Preprint*, arXiv:2302.12039.

Rishabh Kaushal, Jacob van de Kerkhof, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2024. Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT '24)*, page 19 pages, Rio de Janeiro, Brazil. ACM.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-Mod: Can Large Language Models Assist Content Moderation? In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. *arXiv preprint arXiv:2309.14517v2*.

Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6.

Chun-Hsien Lin and Pu-Jen Cheng. 2024. Legal documents drafting with fine-tuned pre-trained large language model. *Preprint*, arXiv:2406.04202.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. 2024. Better Call GPT, Comparing Large Language Models Against Lawyers. *Preprint*, arXiv:2401.16212.

M. E. J. Masson and M. A. Waldron. 1994. Comprehension of legal contracts by non-experts: Effectiveness of plain language redrafting. *Applied Cognitive Psychology*, 8:67–85. P. 16.

Peter Melinat, Tolja Kreuzkam, and Dirk Stamer. 2014. Information overload: A systematic literature review. `https://doi.org/10.13140/2.1.4293.7606`.

Jiayi Pan, Chengcan Wang, Kaifu Zheng, Yangguang Li, Zhenyu Wang, and Bin Feng. 2023. Smoothquant+: Accurate and efficient 4-bit post-training weightquantization for llm. *Preprint*, arXiv:2312.03788.

Weicong Qin and Zhongxiang Sun. 2024. Exploring the nexus of large language models and legal systems: A short survey. *Preprint*, arXiv:2404.00990.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Antonino Rotolo and Giovanni Sartor. 2023. Argumentation and explanation in the law. *Frontiers in Artificial Intelligence*, 6.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *Preprint*, arXiv:2402.07927.

Heiko Thimm. 2023. Data modeling and NLP-based scoring method to assess the relevance of environmental regulatory announcements. *Environmental Systems and Decisions*, 43:416–432.

Amaury Trujillo, Tiziano Fagni, and Stefano Cresci. 2024. The DSA Transparency Database: Auditing self-reported moderation actions by social media. *arXiv preprint arXiv:2312.10269*.

Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial intelligence*, 291:103404.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024. Unsupervised information refinement training of large language models for retrieval-augmented generation. *Preprint*, arXiv:2402.18150.

Xiaoxian Yang, Zhifeng Wang, Qi Wang, Ke Wei, Kaiqi Zhang, and Jiangang Shi. 2024. Large language models for automated Q&A involving legal documents: a survey on algorithms, frameworks and applications. *International Journal of Web Information Systems*, 20(4):413–435.

Yu-Cheng Zhou, Zhe Zheng, Jia-Rui Lin, and Xin-Zheng Lu. 2022. Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking. *Computers in Industry*, 142:103746.