"You Are Orthos": An Experiment in AI-Assisted Human-Rights-Based Reasoning for Content Moderation Disputes under the Digital Services Act

Lorenzo Gradoni[a], Matteo Magnini[b]

[a]*Luxembourg Center for European Law, University of Luxembourg, Weicker 4, rue Alphonse Weicker, Luxembourg, 2721, Luxembourg*
[b]*Department of Computer Science, University of Luxembourg, Maison du Nombre, 6, avenue de la Fonte, Esch sur Alzette, 4364, Luxembourg*

Abstract

This paper investigates whether contemporary Artificial Intelligence (AI) systems can assist the resolution of online content moderation disputes within the institutional framework created by Article 21 of the Digital Services Act (DSA), which enables the establishment of certified out-of-court dispute settlement bodies. It presents Orthos, an AI-assisted decision-support system designed to generate human rights-based legal reasoning for disputes between users and platforms. The paper reports the results of an experiment in which the system was applied to a dataset of moderation cases, with the aim of assessing the capacity of state-of-the-art Large Language Models (LLMs) to support human rights-based adjudication at scale. Its central question is whether contemporary LLMs can generate forms of legal reasoning sufficiently disciplined and coherent to assist human rights-based adjudication in high-volume moderation disputes.

Keywords: Digital Services Act, Content Moderation, Human Rights, Artificial Intelligence, Large Language Models, Legal Reasoning

1. Introduction

This paper examines whether contemporary Artificial Intelligence (AI) systems can assist in resolving online content moderation disputes within the institutional framework created by Article 21 of the Digital Services Act (DSA) [1, 2]. Since 2024, this framework enables the establishment and operation of certified out-of-court dispute settlement (ODS) bodies across Europe [3, 4, 5]. At the core of this framework lies a problem of scale: disputes are generated in massive quantities, while their resolution requires forms of reasoning that resist automation. Against this background, the paper presents Orthos, an AI-assisted decision-support system designed to generate human rights-based legal reasoning for disputes between users and platforms. It reports the results of an experiment in which the system was applied to a dataset of moderation cases, with the aim of assessing the capacity of state-of-the-art Large Language Models (LLMs) to support human rights-based adjudication at scale.

Online platforms have become central infrastructures for public communication. Social networks, video-sharing services, and discussion platforms host vast quantities of user-generated expression, and increasingly function as primary venues for political debate and everyday interaction [6]. The governance of these environments depends heavily on content moderation, the rules-based practices through which platforms remove, restrict, or otherwise regulate user content [7].

The scale of these operations is unprecedented. Platforms adopt billions of moderation decisions every year, ranging from the removal of individual posts to the sus-

Email addresses: lorenzo.gradoni@uni.lu (Lorenzo Gradoni), matteo.magnini@uni.lu (Matteo Magnini)

URL: https://orcid.org/0009-0001-7189-1838 (Lorenzo Gradoni), https://orcid.org/0000-0001-9990-420X (Matteo Magnini)

pension or termination of user accounts. While many decisions concern clearly harmful material, others involve expression whose status under freedom of expression principles is genuinely uncertain. Content moderation disputes typically reflect tensions between competing values: on one side lies the user's freedom of expression and interest in participating in digital life; on the other lies the platform's responsibility to enforce community standards and protect users from harm. Assessing the legitimacy of moderation measures therefore often requires contextual judgment and balancing [8].

As the volume of online expression continues to grow, so too does the number of disputes arising from moderation decisions. Mechanisms capable of reviewing platform actions must be both procedurally fair and capable of operating at high volume. Traditional judicial mechanisms are poorly suited to resolving such disputes at scale, while internal complaint systems operated by platforms often struggle to provide transparent and principled reasoning. Article 21 of the DSA responds to this challenge by enabling the establishment of certified ODS bodies across Europe [9].

The remainder of the paper proceeds as follows. Section 2 examines the institutional framework of Article 21 of the DSA and the practical challenges of adjudicating moderation disputes at scale. Section 3 develops the design of the Orthos system, outlining its core components: a human-rights-based analytical framework centred on proportionality, the generation of bifurcated (user- and platform-protective) reasoned rulings, and the human–machine interaction protocol through which competing lines of reasoning are explored. Section 4 then puts this design to the test, presenting the experimental setup and evaluation framework through which the system's performance is assessed. Section 5 examines the results of the experiment from the perspective of the scoring protocol, assessing the system's overall behaviour across the full set of rulings. Section 6 turns to selected cases that illuminate specific methodological or doctrinal questions, reveal limitations or boundary conditions of the system, and briefly discuss possible fixes. Section 7 concludes.

2. Background

Article 21 of the DSA introduces a system of certified ODS bodies tasked with reviewing moderation decisions taken by online platforms [1, 3]. Users who disagree with a platform's action – whether the removal of content, the suspension of an account, or the refusal to remove allegedly harmful material – may submit the dispute to a certified ODS body, typically after exhausting the platform's internal complaint procedure. These bodies operate independently of platforms, and provide an additional avenue of redress alongside internal complaint systems and judicial remedies. Their decisions are not binding, a feature that may itself prove advantageous. By lowering the stakes of participation, non-binding review may facilitate cooperation. This cooperative logic is reinforced by the DSA's economic design, under which platforms must bear the costs of the procedure and therefore have, in principle, an interest in the smooth functioning of the system.

The emergence of this mechanism should be understood against the background of earlier attempts to subject platform moderation decisions to quasi-judicial oversight. A significant milestone in this evolution was the creation of the Meta Oversight Board in 2020 [10, 11]. Operating through a trust structure in which Meta funds the institution through periodic irrevocable tranches, the Board was designed as an independent body reviewing a limited number of emblematic moderation disputes in light of human rights standards, which occupy a prominent place in its jurisprudence. Its jurisdiction, however, remains limited to Meta's platforms. The initiative showed that moderation decisions could be examined through procedures resembling judicial reasoning, but its architecture – based on discretionary case selection and operating at considerable cost – was not designed to address disputes at scale.

Article 21 can be read as an attempt to generalise this emerging model of review while embedding it within a public regulatory framework. Instead of a single corporate body exercising selective oversight, the DSA creates the conditions for an ecosystem of certified dispute settlement bodies capable of reviewing moderation decisions across platforms and at scale. Certification is granted by a national Digital Services Coordinator and may enable a body to process cases from anywhere in the EU. Article 21 thus establishes a transnational institutional field shaped by the experimentalist logic characteristic of recent European digital regulation: rather than prescribing a single institutional template, the DSA creates a framework within which multiple bodies may operate, each potentially developing its own operational strategies, proce-

dural arrangements, and interpretative practices.

The economic design of this system is equally distinctive. Platforms must bear the costs of dispute settlement, while users must be able to access the procedure free of charge or for a nominal fee. The fees charged to platforms may not exceed the costs incurred by the dispute settlement body. This arrangement excludes profit-maximising business models and instead creates a cost-recovery framework in which the activity of ODS bodies resembles the provision of a regulated public service rather than a commercial arbitration market. From an institutional perspective, the mechanism accordingly functions as a form of delegated justice financed by the platforms whose decisions are subject to review [9, 12].

The institutional design of Article 21 of the DSA is premised on the objective of resolving content moderation disputes at scale. Online platforms process vast volumes of user-generated content, and moderation decisions are adopted on a comparable scale. Public transparency data indicate that platforms adopt billions of moderation decisions annually, each of which could in principle give rise to a complaint. Even if only a very small fraction of these decisions were challenged, the number of disputes potentially reaching ODS bodies would still be considerable. A tentative estimate based on the caseload history of Meta's Oversight Board suggests that a mature Article 21 ecosystem could plausibly attract hundreds of thousands of disputes per year, although the exact scale remains difficult to estimate.[1]

Settling each of these disputes calls for contextual legal reasoning about the meaning and circumstances of the contested expression. It also requires an assessment of whether the moderation measure – or the decision not to adopt one – struck an appropriate balance between competing interests, most notably the user's freedom of expression, the fundamental rights of others, and the platform's responsibility to maintain a safe and functional communication environment.

Producing such reasoning consistently across large numbers of cases presents a formidable challenge. Human adjudication is necessarily time- and resource-intensive, while purely automated moderation systems are ill-suited to contextual legal reasoning. The effectiveness of the Article 21 mechanism therefore depends on the development of approaches capable of sustaining adjudication at scale while preserving the normative and rhetorical discipline expected in disputes concerning freedom of expression.

This tension between scale and legitimacy provides the point of departure for the system presented in the next section. Rather than attempting to replace human judgment, the approach developed here seeks to support human-rights-based adjudication by structuring and amplifying legal reasoning in a form capable of operating under conditions of volume. The following section presents Orthos as a concrete attempt to meet that challenge.

3. Orthos: System Design

One possible response to the challenge of adjudicating moderation disputes at scale is through AI systems capable of supporting the production of legal reasoning [13]. Orthos is designed as a decision-support system, aimed at assisting human adjudicators. Its core function is to generate alternative, human-rights-based reasoned rulings, enabling decision-makers to weigh competing interpretations of a case. Figure 1 illustrates the overall framework. A case concerning a user's challenge to a platform moderation decision is input into the system. Orthos[2] takes the factual scenario and the additional legal documents.

[1] A rough estimate places the potential annual volume of disputes at around 1.6 million. This figure is obtained through a simple extrapolation from the publicly reported caseload history of Meta's Oversight Board. At its peak, the Board received roughly 480,000 submissions in a single quarter, a figure that provides a useful reference point because complaint volumes declined once users realised that the Board reviews only a tiny fraction of submissions through a certiorari-style selection process rather than adjudicating disputes at scale. If that level of engagement were sustained across a full year, which is not unlikely, it would correspond to approximately 1.9 million submissions globally. Adjusting this figure to reflect the share of complaints originating in Europe – about one quarter of the total – yields roughly 475,000 potential European disputes annually relating to Meta platforms alone. Since ODS bodies under Article 21 may in principle review disputes involving several major very large online platforms (including services such as TikTok, YouTube, or X), scaling the estimate to reflect the wider platform ecosystem produces an order-of-magnitude figure of roughly 1.6 million possible disputes per year in Europe. This calculation is necessarily speculative and intended only to illustrate the potential depth of the dispute-resolution field rather than to predict actual case volumes.

[2] The two-headed dog of Greek mythology.

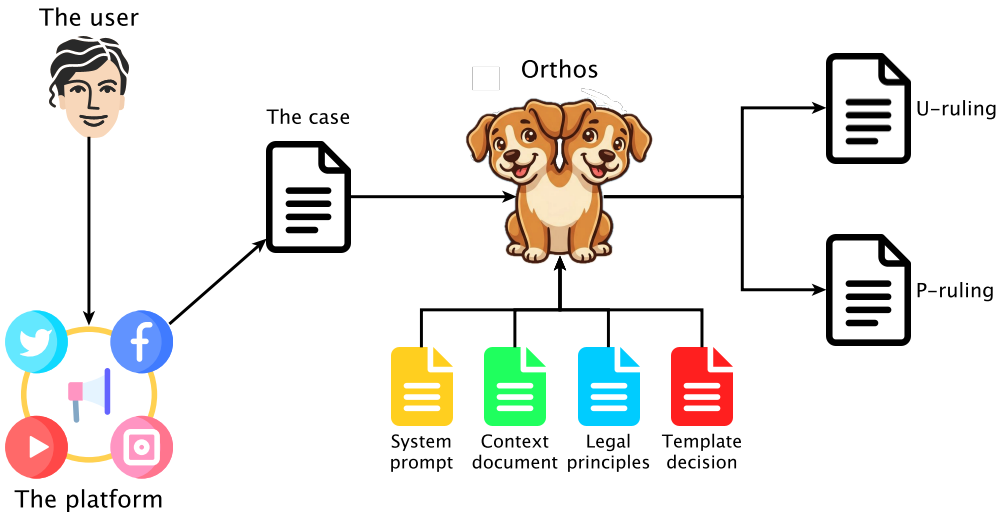

Figure 1: The architecture of Orthos, an AI-assisted decision-support system designed to generate human rights-based legal reasoning for content moderation disputes under the Digital Services Act. The system produces two parallel rulings for each case: one favouring the user (U-ruling) and one favouring the platform (P-ruling), both structured according to the stages of proportionality analysis.

Orthos operates through a bifurcated architecture: for every dispute, it produces two draft rulings – one favouring the user (U-ruling), one favouring the platform (P-ruling) – relying on the same facts and legal framework. Both rulings apply proportionality analysis, examining if the moderation measure pursued a legitimate aim, was suitable and necessary, and balanced interests appropriately [14, 15]. This format showcases the contestable nature of moderation: questions on context, expression, proportionality rarely admit a single self-evident answer. By developing parallel reasoning, Orthos makes the normative tensions explicit, inviting adjudicators to judge which interpretation is most persuasive. In this sense, Orthos is a rhetorical machine – not an outcome calculator, but a generator of legal arguments [13].

The experiment of this paper asks whether contemporary LLMs can produce reasoning with enough coherence and discipline to genuinely assist adjudication in high-volume environments [16, 17, 18]. To do so, the system must rest on solid normative ground. Generating alternative rulings presupposes a choice about the legal language: in which framework will moderation disputes be articulated and assessed?

Moderation disputes arise within a layered normative environment combining contractual rules, domestic legal systems, and human rights principles. Determining how these sources interact is crucial for any dispute settlement body expected to operate across platforms and jurisdictions, yet Article 21 of the DSA itself does not resolve this question. At first sight, the answer might appear straightforward. Most moderation decisions implement platform terms of service, which constitute the contractual framework governing the relationship between platforms and users. From this perspective, an ODS body would simply assess whether the platform correctly applied its own rules. Yet such an approach risks reaffirming the dominant position platforms already occupy in defining and enforcing the norms governing online speech—a situation the DSA seeks to mitigate. If dispute settlement were confined to verifying the internal logic of platform policies, the procedure would offer little more than an externalised version of the platform's own complaint-handling system. A second possibility would be to treat moderation disputes as ordinary legal conflicts governed by national law, including constitutional and statutory protections of freedom of expression. However, this solution proves difficult to operationalise in a transnational setting. Platforms operate simultaneously across multiple ju-

risdictions, while ODS bodies may review disputes originating from users located anywhere in the EU. Applying domestic legal regimes on a case-by-case basis would introduce a level of complexity capable of undermining the accessibility and speed the system is intended to provide.

A third route is to treat human rights law as the main normative language [19]. Within the European legal space, freedom of expression as protected by the Charter of Fundamental Rights of the European Union and the European Convention on Human Rights provides a widely recognised normative anchor for evaluating restrictions on speech. [20, 21]. In this respect, a human-rights-based approach performs a dual function: first, it operates as a complexity-reduction device, allowing adjudicators to analyse disputes through a common normative framework rather than navigating the full diversity of national legal systems; second, it enhances the legitimacy of the dispute settlement process by grounding moderation review in principles that transcend both the contractual preferences of individual platforms and the fragmentation of domestic legal systems.

But even with human rights as a benchmark, the operational challenge remains: how to deliver disciplined, consistent reasoning at scale, especially amid high volumes and diverse controversies? This calls for a closer look at the logic of human rights adjudication itself.

Human rights adjudication is often framed as the ultimate domain of human judgment resisting computation [22]. Human rights-based reasoning unfolds through carefully crafted narratives embedded in analytical frameworks such as the multi-stage proportionality test. These narratives render the exercise of judgment intelligible and publicly justifiable without eliminating indeterminacy. Proportionality analysis illustrates this dynamic particularly well. Courts typically proceed through a sequence of analytical stages, each of which invites the articulation of arguments about the relationship between the contested measure and the interests it seeks to protect or impinges upon. While these steps impose a recognizable structure on the reasoning process, they do not determine the outcome; they just map the argumentative space within which the decision emerges. LLMs are particularly well suited to operate in this rhetorical space. By identifying patterns across large textual corpora, such models can reproduce the argumentative forms characteristic of legal reasoning, including the narrative style of proportionality analysis [13].

Much of the literature on AI in law echoes this middle path. Early symbolic or rule-based systems worked well in formal domains, but struggled where context mattered. Recent Machine Learning (ML) and Natural Language Processing (NLP) advances have revived interest in AI-assisted reasoning. But full automation is controversial: opacity, lack of countability all loom large [23, 24]. Current practice leans toward technological assistance – systems help with research, document analysis, triage – rather than outright replacing the adjudicator. Orthos builds on this trend. Rather than predicting the outcome or optimizing case management, it focuses on disciplined generation of legal reasoning: mapping plausible arguments for both sides and supporting human adjudicators in complex scenarios.

The practical feasibility of AI-assisted reasoning is probed here through two experiments: an exploratory exercise with a synthetic case, and a systematic study on a larger dataset of real moderation disputes – the empirical base of this article.

4. Experiment

We designed a two-stage experimental protocol to assess whether contemporary LLMs can generate human rights-based legal reasoning for platform moderation disputes under Article 21 of the DSA.

The data, code and detailed configuration are available at `https://github.com/MatteoMagnini/experiments-rulings-cases-with-llm`.

4.1. Stage 1: Human-Chatbot Dialogue

The first stage centered on a direct conversation between a domain expert and a ChatGPT-5 chatbot instructed to impersonate Orthos. The scenario simulates a typical dispute: the author of a social-media post challenges its removal by a platform. The chatbot, equipped with the Orthos prompting environment, is tasked with generating two reasoned rulings – one in favour of the user (U-ruling), one in favour of the platform (P-ruling) – each structured according to the stages of proportionality analysis: legitimate aim, suitability, necessity, and balancing, following the framework articulated by Robert Alexy [14, 15] among others. In this interaction, the

model demonstrates a strong command of the theory and vocabulary of proportionality analysis. However, the experiment reveals critical weaknesses. In the U-ruling, where the moderation measure was found to fail the suitability requirement, the model was prompted to reason conditionally through the remaining stages of necessity and balancing to reinforce the conclusion. When generating the corresponding P-ruling, however, the model mistakenly preserved this conditional cascade, overlooking that a moderation measure must satisfy all three requirements to be deemed a legitimate restriction of speech. The model thus failed to grasp the asymmetry between the U-ruling's one-strike-out logic and the P-ruling's requirement that all three proportionality conditions be satisfied. This results in outputs that imitate the structure of legal analysis while misconstruing its logic—a phenomenon that falls under the category known as the "Potemkin effect" [25, 13].

4.2. Stage 2: Systematic Dataset Evaluation

The second stage expands the approach to a curated dataset of 45 real-world moderation disputes. The objective is to test the LLM's capacity to produce logically disciplined and persuasive legal argumentation across cases that differ in type and complexity. In this experiment we used GPT-5.2, with a temperature of 0.1 and maximum number of output tokens set to 4,096. Having a low temperature but higher than 0 intuitively allows the model to stay close to the facts and increase its storytelling capabilities, which are both important for the task at hand. We detail the dataset composition and prompt structure below.

4.2.1. Dataset

We collected 45 cases published by *User Rights*, an ODS certified by the German Digital Services Coordinator in August 2024. At the time of writing, User Right's is the only accessible corpus of decisions issued under Article 21 of the DSA [26]. Each case outlines a factual scenario involving a major platform: content removals, refusals to act, or account suspensions. Descriptions are extracted and standardized via NotebookLM, yielding concise factual statements capturing the essentials of each dispute. The dataset encompasses key scenarios of contemporary moderation: alleged hate speech, harassment, misinformation, artistic expression, extremist symbolism, impersonation; both takedown and leave-up cases are present. This range allows us to test the system's reasoning across a wide array of factual and legal contexts.

4.2.2. Prompt Suite

The experiment relied on a four-component prompt suite used to simulate the reasoning environment of an ODS body operating under Article 21 of the DSA. The suite consisted of four text files providing complementary instructions and contextual material, organised from general orientation to concrete output format: (1) System Prompt; (2) Context Document; (3) Legal Principles: Freedom of Expression and Conditions Under Which It May Be Restricted; (4) Template Decision.

System Prompt. The system prompt opens with the instruction *"You are Orthos, a rhetorical machine designed to assist in the resolution of disputes concerning online content moderation under Article 21 of Regulation (EU) 2022/2065 (Digital Services Act)."* This opening line establishes the institutional role of the system and the adjudicative setting in which it operates. The prompt immediately clarifies that the system does not decide cases but assists a human decision-maker by generating alternative reasoned outcomes: *"For every case, you must produce two outputs: a ruling in favour of the user (U-Ruling) and a ruling in favour of the platform (P-Ruling)."* The assigned task is therefore argumentative rather than predictive: the model is instructed to trace two legally plausible lines of reasoning based on the same factual record.

The core analytical instruction concerns the structure of the proportionality test. Each ruling must examine whether the platform's intervention pursues a legitimate aim, whether the measure is suitable, whether it is necessary, and whether it strikes a permissible balance between the competing interests at stake. The prompt explicitly frames this analysis within human-rights reasoning, referring in particular to Article 11 of the Charter of Fundamental Rights of the European Union [20], Article 10 of the European Convention on Human Rights [21].

A distinctive feature of the prompt is that it instructs the model to identify the procedural posture of the dispute before applying the proportionality analysis, since proportionality operates differently depending on whether the dispute concerns the removal or the non-removal of content. The prompt distinguishes between takedown cases,

where a platform removed or restricted content, and leave-up cases, where the content remained online despite a complaint. The structure of the reasoning differs depending on the type of dispute. In takedown cases, the user-protective ruling follows a "one-strike-out" logic, meaning that the restriction is deemed disproportionate if it fails at any stage of the proportionality test. In leave-up cases, the logic is reversed: the ruling favouring removal must satisfy all stages of the proportionality analysis, while the ruling favouring the platform may rely on failure at any single stage to justify leaving the content online.

Finally, the prompt imposes rhetorical constraints on the outputs. The rulings must follow the structure of the template decision provided in the prompt suite and must read as "reasoned decisions in an out-of-court dispute settlement procedure," written in clear and accessible language but maintaining the rhetoric of legal reasoning.

Context Document. The context document provides a conceptual description of the Orthos system and the institutional setting in which it operates, that is, out-of-court dispute settlement under Article 21 of the DSA. The document also situates moderation disputes within the particular environment of online platforms, which simultaneously host large-scale public discourse and enforce their own community standards.

Legal Principles Document. The third component of the prompt suite provided a concise statement of freedom-of-expression principles drawn primarily from the Council of Europe's Guide on Article 10 of the European Convention on Human Rights [27], which served as the main doctrinal source for the prompt. Rather than reproducing doctrine in detail, the prompt condensed key elements of Article 10 case-law into a short operational summary designed to guide the model's reasoning. In compressed form, the document conveyed four core ideas.

First, freedom of expression is described as a foundational principle of a democratic society, protecting not only agreeable speech but also expression that may "offend, shock or disturb". Second, the document explains that restrictions on expression may be justified when necessary to protect legitimate interests such as the rights and dignity of others, public safety, or the integrity of online communities. Third, the text emphasises that disputes involving platforms take place in a "horizontal" setting: the conflict is not between an individual and the State but between a user and a private intermediary that governs a digital communication environment. The assessment must therefore balance the platform's regulatory autonomy with the user's interest in participating in public discourse. Finally, the document highlights several contextual factors typically used in Article 10 analysis, including the subject matter of the speech, the role of the speaker, the status of the person targeted, the form and tone of the expression – for example satire, artistic expression, or political criticism – and the potential impact of the speech on others.

The purpose of this document was therefore not to reproduce the full complexity of freedom-of-expression doctrine – an undertaking that would in any event have been impracticable within a prompt – but to provide a set of condensed doctrinal building blocks capable of guiding proportionality reasoning while remaining flexible enough to be applied to new configurations of facts, different from the one underlying the model decision that constitutes the fourth and final component of the prompt.

Template Decision. The final component of the prompt suite defines the structure and style expected of the model's outputs. It reproduces a template decision – *Amina K. v. TikTok* – illustrating the canonical format of an Orthos ruling. The *Amina K.* case is the synthetic dispute used in the first Orthos experiment as a controlled test case for developing the ruling structure [13].

The template required each ruling to follow a three-part structure: (1) The Case, presenting the factual background; (2) Assessment, applying the proportionality analysis; (3) Decision, stating the outcome. Both rulings – the U-Ruling and the P-Ruling – were required to contain an identical factual section and to proceed through the same analytical steps, ensuring structural symmetry between the two outcomes.

Importantly, the template decision was formulated as a takedown case – that is, a situation in which content has been removed – the most common form of moderation decision, though not the only one. Confining the prompt suite to a single template was a deliberate methodological choice. It allowed us to examine whether the system could transpose the template to other configurations

Table 1: Evaluation metrics for Orthos' generated rulings. For each metric, the table reports the mean and standard deviation across the 45 cases. The p-value column reports the p-value of the Mann-Whitney U statistical test between the U-ruling and P-ruling metrics. The null hypothesis is refused on the persuasiveness metric (statistically significant group means in bold).

Metric	U-ruling	P-ruling	p-value
Fidelity	4.96 ± 0.21	4.58 ± 1.10	0.070
Context	4.62 ± 0.81	4.44 ± 1.10	0.684
Proportionality	3.84 ± 0.37	3.76 ± 0.74	0.649
Persuasiveness	**3.91 ± 0.51**	3.51 ± 0.82	0.008
Institutional	3.82 ± 0.58	3.69 ± 0.63	0.166
Metric	Pair		
Symmetry	4.44 ± 1.16		
Plausibility	4.22 ± 0.88		
Global	4.07 ± 1.03		

of dispute, in particular leave-up cases and disputes involving account suspensions, even though these scenarios require – as we shall see in Section 6 – partially different analytical approaches and reasoning patterns.

5. Results

The evaluation of the rulings generated in the experiment relied on eight analytical criteria aimed at capturing the main dimensions of adjudicative quality. These criteria assess the extent to which the system faithfully reconstructs the facts, correctly interprets the communicative context of the expression, meaningfully applies proportionality reasoning, and produces persuasive lines of reasoning. They further examine structural features specific to the Orthos architecture, including the symmetry between the two alternative rulings and the degree to which the decisions adopt an institutional tone appropriate for dispute settlement under Article 21 DSA. Finally, the evaluation considers whether the pair of rulings appears as a plausible set of legal artifacts and includes a global score capturing the evaluator's holistic assessment. All ratings were assigned by one of the authors on a 1-5 Likert scale.

Table 1 summarises the results: means and standard deviations for each metric, plus the p-value of the Mann-Whitney U test comparing U- and P-rulings. We can already appreciate that in general the model performs quite well in all the dimensions.

5.1. Fidelity to the Facts

This dimension was assessed using the following question: "Does the ruling accurately represent the facts of the case without omission, distortion, or invention?". We assessed this by examining whether the ruling faithfully reconstructs the relevant events, actors, and procedural posture based on the case description – the only case-specific input – without omitting relevant details, distorting the narrative, or introducing elements not contained in the file. Since Orthos generates the factual section and reasoning probabilistically rather than reproducing a fixed template, particular attention was paid to the system's ability to avoid hallucination and maintain strict fidelity to the factual record.

The system performed very strongly on this parameter, as reflected in the scores. In most cases, both the U- and P-rulings received the maximum score for factual fidelity, indicating that the model almost always reproduced the case narrative accurately and avoided introducing invented elements. Only a small number of rulings showed distortions of the input facts, typically affecting one side of the pair, most often the platform-protective ruling. These distortions appear to arise when the model searches for argumentative footholds to defend positions at the outer edge of plausibility. They are readily recognisable as elements of strained reasoning and carry little persuasive force.

This strong performance comes with a trade-off. The model's fidelity to the factual input sometimes appeared overly rigid: Orthos adheres closely to the wording and scope of the case description and never extends the narrative beyond what is explicitly provided. Within these constraints, however, the template-based subdivision of the factual section into short paragraphs contributes to clear and readable case descriptions.

A key aspect of factual fidelity is the system's ability to reason under conditions of evidentiary incompleteness. Instead of filling gaps in the record with invented facts, the rulings sometimes acknowledge the limits of the available information and treat them as a parameter in the analysis. This approach is illustrated in the following passage from the *Gestohlene Identität* case: *"[t]he facts provided*

do not specify whether those threats were made through the account itself, through private messages, or through another channel, nor do they specify what information Instagram had before it when it decided not to suspend [...]. [A] platform may reasonably require a clear link between the reported account and the alleged threats, or sufficient information to act confidently, before applying an account-level description" (UR-2025-24, P-ruling, para. 11).

The only clear instance of hallucination occurs in the *Nicht Verschreibungspflichtig* case, where it is stated that *"[t]he complainant argues, in substance, that paracetamol is commonly available without prescription"* (UR-2026-10, P-ruling, para. 10). The claim about the medication is correct, but the statement of facts does not indicate that this position was advanced by the user, nor does it suggest otherwise. The inference is plausible, since the argument weighs in favour of the user, but it nonetheless remains unsupported by the record. This minor misstep is ultimately harmless, as arguments supporting either position are typically raised *ex officio* in a human rights-oriented adjudicative context.

5.2. Contextual Understanding

This aspect of the analysis turned on the following question: "Does the ruling correctly interpret the nature and context of the expression?". Effective reasoning depends on situating expressive acts within their communicative setting (politics, satire, artistic expression, harassment, meme, etc.), tracking cues provided by the conversation, speaker intent, and platform-specific context. Scores on this parameter are consistently high, typically falling between 4 – good contextual awareness – and 5 – excellent nuanced analysis –, with little imbalance between U- and P-rulings. The following examples illustrate how this performance manifests across different types of context-sensitive cases.

Orthos shows a consistent ability to recognise culturally coded signals embedded in the material, as illustrated inter alia in the *Verbotene Ideologien* case (UR-2026-08), which turns on references to numerical symbols associated with extremist ideologies such as "14" or "88" (whose meaning was neither explained in the input nor known to the evaluator). A similar sensitivity to coded context appears in the *Sylt-Kommentar* case (UR-2025-08), where the comment "the song was improved by Sylt" was recognised as alluding to a widely reported incident in which racist chants were sung to the melody of *L'amour toujours* on the German island of Sylt, now circulating online as a meme associated with xenophobic signalling.

A different form of contextual sensitivity appears in the *Banned for Bullying* case (UR-2025-21), which turns on the reconstruction of a conversational setting. An insulting expression was posted under a thread already containing aggressive language associated with a pre-fight exchange between mixed martial arts fighters. In that setting, expressions such as the disputed insult and similar forms of verbal provocation belong to a well-established promotional style often described as "trash talk," which dramatizes competition rather than initiating personal harassment. The user's repetition of the term thus appeared embedded in a confrontational rhetorical environment rather than introducing a new form of hostility into the discussion. The U-ruling identifies this contextual feature and treats it as relevant to the proportionality assessment of the platform's response. By situating the expression within the dynamics of the surrounding exchange rather than evaluating it in isolation, the ruling demonstrates a capacity to reconstruct the relevant communicative setting.

Another aspect of contextual understanding lies in the system's ability to navigate competing registers of meaning, using the tension between literal and metaphorical readings as an argumentative lever. In the *Gewaltaufruf* case, the U-ruling states that saying that the "head" of a public figure "must roll" can be understood *"as a call for violence, but it is also a well-known idiom sometimes used as political hyperbole to express anger or a demand for accountability"* (UR-2025-23, U-ruling, para. 8). By contrast, the P-ruling shifts the focus to the operational constraints of platforms, noting that they operate *"at scale and must often act on the basis of the content's plain meaning"* and that *"[h]ere, the plain meaning is violent"* (para. 8).

The system also performs well in cases involving composite forms of expression, where multiple communicative strands coexist within a single piece of content. In the *Corona Fehlinformationen* case, the ruling recognises that *"[t]he video [...] combined political speech (an interview with a prominent politician), commentary by a public figure advancing conspiratorial narratives, and the com-*

plainant's personal account of alleged vaccine injury" (UR-2025-06, U-ruling, para. 10). Rather than collapsing these elements into a single category, the reasoning disaggregates them and draws differentiated inferences from each component, allowing the proportionality analysis to track their distinct normative valences. This capacity to parse layered messages is particularly valuable in content moderation disputes, where meaning and impact often depend on the interaction between multiple registers.

The system also demonstrates a clear understanding of platform-specific communicative environments, integrating considerations relating to platform design, audience reach, and modes of content circulation into its reasoning, including in user-protective rulings. This sensitivity is evident across several cases. In the *Trump and the Dictators* case, the ruling notes that *"[i]n a fast moving, highly shareable environment, the platform is entitled to consider not only the author's intended meaning but also the foreseeable ways content may be consumed, misunderstood, or weaponised"* (UR-2025-28, P-ruling, para. 9). Similarly, in the *Puppyplay* case, it observes that *"[a] platform may apply stricter standards than those applicable in public space, particularly where its service is used by a wide age range and content is delivered algorithmically to users who did not seek it out"* (UR-2026-07, P-ruling, para. 7). A comparable concern with circulation and audience reception appears in the *AfD-Reden* case, where the ruling, confronted with extremist statements reproduced for the purpose of criticising them, emphasises that *"[i]n short-form video formats, viewers may encounter the statements without absorbing the surrounding framing, and the statements may be clipped, re-uploaded, or shared in ways that detach them from the caption or the concluding call to vote for democratic parties"* (UR-2025-10, P-ruling, paras. 9). This sensitivity extends to platform-specific communicative norms. In the *Protest gegen Rechts* case, the ruling recognises that LinkedIn's professional orientation may justify stricter moderation of political polemic, noting that *"LinkedIn's interest in [...] maintaining a professional communicative environment can justify the interference"* (UR-2025-07, P-ruling, para. 14).

A minor weakness lies in the occasional inclusion of analytically unnecessary disclaimers. In the *Heimliche Aufnahme* case, the ruling states that *"[t]he complainant's expression is not political speech, but it is ordinary social communication and humour"* (UR-2026-06, U-ruling, para. 13), even though the non-political nature of the expression is obvious from the context. The clarification does not advance the reasoning and instead introduces a slightly artificial tone. The same pattern appears in the *Mobbing Minderjähriger* case, where Orthos explicitly notes that insults targeting a child's weight do not amount to *"political speech, journalism, or debate on a matter of public interest"* (UR-2026-13, U-ruling, para. 10).

A final limitation arises where meaning depends on recent events outside the case record. In the *Satire Morbide* case (UR-2025-16), a post showing "Charlie Kirk has started a live video," with imagery suggesting he is in hell, is interpreted as an insult directed at a living figure. The ruling thus misses the decisive contextual element, the sting lying in joking about Mr Kirk's death. The gap reflects missing up-to-date context rather than flawed reasoning. We further discuss the implications of this case in Section 6.1.

5.3. Quality of Proportionality Reasoning

This aspect of the analysis was guided by the following question: "To what extent does the ruling meaningfully apply the quadripartite analysis (legitimate aim, suitability, necessity, and balancing)?". The assessment examined whether these steps are applied meaningfully rather than mechanically, and whether the reasoning clearly links the measure to freedom of expression and the competing rights and interests at stake [14, 15].

A key concern, in light of the previous experiment, was whether Orthos would reproduce the earlier misapplication of the proportionality cascade, correctly treating failure at any stage as decisive in the U-ruling, yet erroneously carrying that logic into the P-ruling, where all requirements must be satisfied (see Section 4.1). The results show that the system consistently avoids this error. Across the dataset, the rulings display a coherent grasp of the asymmetry between the two argumentative paths, applying the quadripartite test in a way that is both logically consistent and attentive to its internal structure.

This strong performance is reflected in the scoring results. Scores on this parameter cluster tightly around 4 – well-developed reasoning – for both U- and P-rulings, with near-perfect symmetry. The scale was applied with a deliberately demanding calibration: a score of 4 reflects a

level of analysis approaching publishable quality, while 5 – sophisticated proportionality analysis – marks an aspirational benchmark for further model refinement. The only noticeable collapse occurs in a single ruling attempting to defend a position that is, on the facts, indefensible. In that instance, the model struggles to identify any meaningful factual foothold on which to articulate the four stages of the analysis, resulting in a visibly weakened reasoning structure. The case is examined in greater detail in Section 6.2.

5.4. Persuasiveness of the Outcome

This dimension was assessed using the following question: "Does the reasoning make the conclusion convincing and defensible, even if one might disagree?". While the preceding section assesses the structure of the analysis, this parameter evaluates its rhetorical force.

Overall performance is strong. Scores cluster around 4 (strong reasoning) for U-rulings, while P-rulings tend to fall slightly above the midpoint between 3 and 4. The scale is deliberately demanding: a score of 4 reflects reasoning approaching publishable quality, while 5 – highly persuasive – sets the horizon for further model development. The gap between user-protective and platform-protective rulings is statistically significant. The asymmetry likely reflects the underlying structure of the cases. In this setting, the user's position is often normatively stronger than the platform's, whose decisions are taken at scale and may rely on standardised responses that fail to take sufficient account of context. It is generally easier to argue a strong case persuasively than to defend a weaker one. From a rhetorical standpoint, the latter remains the more demanding task and marks an area for further improvement.

However, there have been a few instances of excellent defensive manoeuvres in situations of clear argumentative disadvantage. For example, in the *Protest gegen Rechts* case, the P-ruling strengthens the platform's position by deftly adjusting the necessity test, without abandoning the proportionality framework: *"Necessity does not require that removal be the only conceivable response, but that it be a reasonably necessary one in the circumstances"* (UR-2025-07, P-ruling, para. 10). The most compelling performances appear in U-rulings that reframe the platform's own rationale. In the *Suizidandeutung* case, the ruling acknowledges that the user's language is *"alarming"* but emphasises that it may function as *''a cry of help"* or a form of *"seeking support"*, and concludes that removal *"may also undermine the protective aim"* of the policy against suicide and self-harm by cutting off avenues for assistance (UR-2025-25, U-ruling, paras. 6-7). The argument is persuasive precisely because it works from within the platform's protective logic, turning it into a reason against removal.

5.5. Symmetry between U- and P-Rulings

This criterion examines whether the two rulings produced for each case display comparable argumentative quality. It is guided by the following question: "Do both rulings appear equally serious and well reasoned, or is one clearly weaker?". Rather than seeking a single correct outcome, Orthos aims to articulate the strongest plausible case for each side, each time in the form of a ruling rather than a partisan pleading. Symmetry therefore captures the extent to which the pair forms two credible adjudicative paths.

Performance on this parameter is consistently high across the dataset. Scores cluster between 4 – mostly balanced – and 5 – fully symmetrical –, with a clear concentration toward the upper end of the scale. In most cases, both rulings exhibit comparable depth, structure, and rhetorical refinement. Departures from symmetry are rare and occur where one position – almost always the platform's – is, on the facts, hard to defend, leading to strained argumentative moves on one side and a corresponding drop in balance. These instances do not indicate a systematic bias; they reflect the limits of adversarial reconstruction where one side of the dispute lacks a plausible case.

5.6. Institutional Appropriateness

This criterion evaluates whether the rulings adopt a tone and argumentative posture appropriate for decisions issued by an ODS body operating under Article 21 of the DSA. It is guided by the following question: "Does the ruling sound like a measured, non-binding ODS decision, rather than a platform moderation notice or a constitutional court judgment?".

The DSA does not prescribe a drafting style. We adopted, by design, a middle register reflecting the func-

tional position of ODS bodies: decisions should be reasoned and principled yet concise and accessible; authoritative yet explicitly non-binding; attentive to fundamental rights yet responsive to the operational realities of platform governance. This parameter captures the extent to which the rulings approximate this constructed style.

Performance is solid, with U-rulings scoring slightly higher than P-rulings on average. Where scores fall below 4 – good institutional tone –, deviations are minor and take two forms: occasional drift toward doctrinal density, or, conversely, compressed, policy-driven formulations closer to platform notices. These deviations remain limited and do not undermine the overall plausibility of the rulings as outputs of an Article 21 ODS body. As with the previous parameters, a score of 5 – excellent adjudicative style – should be understood as an aspirational benchmark, marking the horizon for further refinement of the model's stylistic calibration.

5.7. Overall Adjudicative Plausibility

This criterion is guided by the following question, shifting the focus from individual rulings to the pair as a whole: "Do the two rulings together read as plausible decisions that a human adjudicator might reasonably reach?". It evaluates whether the two alternative outcomes, taken together, form a credible set of adjudicative possibilities within a human rights framework. In this sense, it captures the overall legal plausibility generated by the interaction of the two argumentative paths.

The average score is 4.22 on a scale where 4 denotes plausible and 5 highly plausible outcomes. In most cases, the two rulings form a coherent and credible set of alternative resolutions that could reasonably be adopted in practice, with only minor refinements.

5.8. Global Assessment

This final parameter provides a holistic appraisal of each pair of rulings, asking: "What is the overall quality of the pair?". Unlike the preceding criteria, which isolate specific dimensions of adjudicative performance, the global score synthesises the evaluator's overall impression after considering the full set of parameters.

Overall performance is strong. Scores sit slightly above 4 on a scale where 4 denotes strong performance and 5 excellent performance. In most cases, the pair of rulings reads as a robust legal artifact, combining fidelity to the record, contextual sensitivity, proportionality reasoning, persuasive presentation of the reasoning, and a style suited to the system's potential institutional role.

6. Discussion

Building on the preceding section's general assessment, this section examines recurring issues that arise in specific types of disputes and reveal the limits of the system's current architecture. They concern, in turn, the occasional dependence of interpretation on contextual information beyond the record; the collapse of symmetry in cases where the correct outcome is self-evident; the limits of a strictly binary remedial structure in leave-up cases; the composite nature of suspension decisions, which call for an intermediate form of ruling; and the risk of overreach where the user-protective ruling introduces measures not sought by the complainant. Taken together, these issues do not undermine the binary architecture of Orthos's rulings; they delineate some of its limits and indicate where it must be supplemented or constrained to accommodate the diversity of content moderation disputes.

6.1. Context Outside the Case Record

The *Satire Morbide* case (UR-2025-16) illustrates a recurring difficulty in AI-assisted adjudication: the interpretation of the contested speech may depend on widely known public events not explicitly mentioned in the case record. The disputed post depicted a notification stating that *"Charlie Kirk has started a live video"*, accompanied by an image of a flaming skeleton in hell. Whether this expression should be interpreted as an insult directed at a living public figure or as mockery following the assassination of Charlie Kirk depends on contextual information not contained in the case record. It is indeed possible that the model's knowledge base at the time of generation did not include the relevant event, highlighting that such interpretive gaps may stem not only from the case record but also from the temporal limits of the model's training data.

One possible approach to addressing this problem would enable the system to incorporate contextual information through a controlled retrieval step. Under this approach, the system would identify "Charlie Kirk" as a named entity and retrieve a short factual note – such as

that Mr. Kirk died following a widely reported shooting in September 2025—which would then be appended to the record before the reasoning process begins.

6.2. Collapse of Symmetry in Easy Cases

The *Foto mit Folgen* case reveals a structural limit of the dual-ruling architecture, one that emerges where there can be little doubt as to what the correct outcome is. The case concerns a story posted by the complainant, a transsexual person, consisting solely of a photo of themselves, with no accompanying text or symbols. The platform removed the content for allegedly violating its policy on violence and dangerous organizations, without providing any explanation. In this instance the takedown appears so clearly unjustified that the symmetry normally expected between the two rulings breaks down. The user-protective ruling, compelled by the system prompt to proceed through all stages of the proportionality analysis, ends up overcomplicating the matter by introducing elements that do not plausibly arise from the facts, for example by suggesting that a warning could constitute a milder alternative even though the content itself does not present a recognizable violation requiring mitigation. What the ruling affirms under the necessity test appears clearly misplaced: *"Less restrictive options – such as leaving the story up, requesting clarification, applying a warning, or conducting a more careful human review – would ordinarily be expected where the content is facially benign"* (UR-2025-18, U-ruling, para. 11). The platform-protective ruling, for its part, struggles to articulate a coherent defense of the platform's action and reads as a painful exercise in strained justification. Rather than undermining the experiment, this type of case is methodologically valuable because it reveals a boundary condition for the dual-ruling architecture. When the illegitimacy of the platform's intervention is overwhelmingly clear, forcing the system to generate a fully developed platform-protective argument is an exercise in futility.

A more appropriate response in such situations is to allow the system to recognize the asymmetry and adapt the output accordingly. Where the platform's action appears manifestly unjustified, the user-protective ruling should be brief and decisive, identifying the absence of any credible ground for interfering with the user's expression. The platform-protective ruling, in turn, need not construct an artificial defense of the decision; it may instead acknowledge the weakness of the platform's position and refer the human decision-maker to the user-protective ruling. Introducing this possibility does not undermine the dual-ruling framework but prevents the system from engaging in performative argumentation where no genuine controversy exists.

6.3. Remedy Structure in Leave-Up Cases

Leave-up cases represented a stress test for the system. As explained in Section 4.2.2, while the system prompt provides specific instructions for such cases, the template decision was deliberately formulated as a takedown scenario. Confining the prompt suite to a single template was intended to assess whether the system could transpose that structure to other configurations of dispute. On this point, the results are convincing: Orthos carries out this transposition in a coherent manner. At the same time, leave-up cases bring into focus a limitation of the current setup.

In leave-up cases, the adjudicator is not reviewing an existing restriction but determining whether a moderation measure should have been adopted. Where the platform has already acted – by removal, demotion, labeling, or otherwise – the dual-ruling structure operates straightforwardly: the P-ruling defends the platform's decision, while the U-ruling advances the user's case. In leave-up cases, by contrast, the decision space is not binary but graduated, with at least three plausible outcomes: maintaining the leave-up decision, ordering removal, or recommending a milder measure.

While complainants often seek takedown, the alleged harm may sometimes be adequately addressed by a less intrusive response, such as contextual labeling or visibility limits. This dilemma cannot be adequately handled by a single user-protective ruling. Since an argument for removal does not proceed in the same way as an argument for a milder measure, forcing both lines of reasoning into one U-ruling would blur the analysis or make it needlessly convoluted, while also depriving the human decision-maker of the opportunity to choose between distinct forms of intervention. The protocol should therefore make this remedial fork explicit through an additional step. It should begin by producing the standard pair of rulings, with the U-ruling making the case for removal, unless the record states that the user requested a milder measure. If the P-ruling persuades, the matter ends there: the

platform's leave-up decision stands and no further inquiry into alternative interventions is required. If the U-ruling persuades, the adjudicator may then request an alternative U-ruling (U-bis) making the case for a less intrusive intervention. In this way, the system preserves the human decision-maker's ability to choose the appropriate intensity of interference with expression.

6.4. Suspensions as a Distinct Sanction

A related but distinct issue arises where the platform's intervention takes the form of account suspension rather than a simple takedown. In such situations, the moderation decision typically comprises two distinct components, though platforms often treat them as a single measure: the removal of the post or posts that triggered the enforcement action, and the suspension of the user's account. Under the basic dual-ruling architecture, the platform-protective ruling validates the suspension while the user-protective ruling concludes that the suspension constitutes an illegitimate restriction on expression. Yet this binary framing obscures an important intermediate possibility. Even where suspension is disproportionate, the underlying post may still violate applicable standards and thus justify removal or a milder intervention. Treating suspension as the sole object of review therefore collapses two distinct questions – the legitimacy of intervening on the content and the proportionality of suspending the account – into a single determination. The *Content 18+* case illustrates the point well. There, Orthos invited the platform to *"(i) reinstate the account; [. . .] and (iii) if it considers specific items to breach its rules, apply a more targeted measure (such as removal of the post or [other] proportionate restrictions) with a clear explanation linked to the relevant policy criteria"* (UR-2025-26, U-ruling, para. 18). This is unsatisfactory, because it leaves unresolved the distinct question whether, and in what form, intervention against the individual piece of content is justified.

To address this difficulty, an adjustment of the protocol is required for cases involving account suspension. In such cases, the system should generate three rulings: a user-protective ruling (U), a platform-protective ruling (P), and an intermediate ruling (M) that quashes suspension while still treating the triggering content as problematic. The U-ruling would find both the takedown and, a fortiori, the suspension disproportionate. The P-ruling would defend suspension. The M-ruling would occupy the intermediate position that the current protocol captures only imperfectly: that the content may legitimately be taken down while suspension remains disproportionate.

The introduction of the M-ruling brings into view a further question, analogous to the one identified in the preceding section. While the M-ruling by definition supports takedown of the specific content, it remains open whether removal is the appropriate response or whether a milder intervention would suffice. As in leave-up cases, this dilemma cannot be resolved within a single ruling without either collapsing distinct lines of reasoning or depriving the human decision-maker of choice among levels of intervention. The protocol should therefore be extended in analogous fashion. In addition to the initial M-ruling, the system should allow the decision-maker to request an M-bis ruling making the case for a milder intervention. This extension preserves the internal symmetry of the architecture while ensuring that sanction calibration – already implicit in the logic of the M-ruling – can be carried through to a final determination that fully settles the dispute rather than leaving unresolved issues liable to generate a further dispute. Figure 2 shows the extended architecture for leave-up and suspension cases, incorporating the additional rulings and the possibility of requesting alternative rulings where appropriate.

6.5. Overreach in Leave-Up Cases

A distinct procedural issue arises in leave-up cases when the user-protective ruling exceeds the scope of the dispute by recommending interventions not sought by the complainant. This tendency is illustrated by the *(No) Time to Say Goodbye* case, where the U-ruling invites the platform not only to take down the content in dispute, but also to *"assess"* whether the author *"should also be subject to account-level measures"* (UR-2026-01, U-ruling, para. 17). This amounts to an unwarranted escalation. Unless the complainant has expressly sought account suspension – as may occur in cases of hostile impersonation or comparable abuses – the U-ruling should not contemplate such measure. The system should therefore be designed to recognize and enforce this boundary, so as to preserve procedural fairness and the integrity of the adjudicative process.

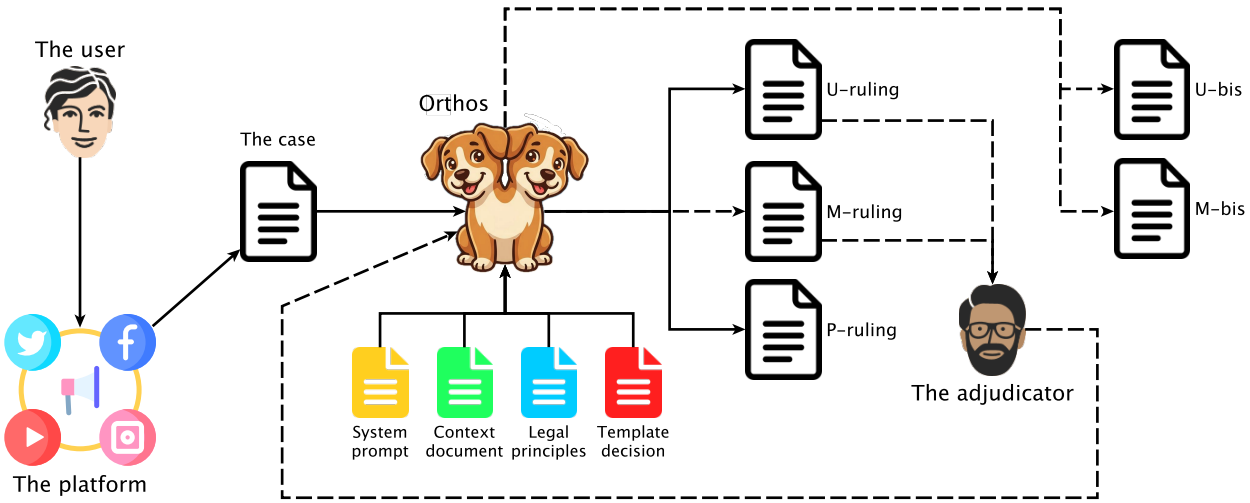

Figure 2: Proposed architecture for leave-up and suspension cases, incorporating additional rulings and the possibility of requesting alternative rulings where appropriate. The adjudicator can ask Orthos to generate U-bis and M-bis rulings to explore alternative interventions where the initial U- or M-ruling supports a more intrusive measure than necessary.

7. Conclusion

The present study demonstrates that contemporary LLMs, when guided by a disciplined prompting architecture, can reliably generate reasoned legal argumentation for high-volume content moderation disputes under Article 21 of the DSA. Orthos displays robust performance across factual fidelity, contextual awareness, proportionality analysis, persuasive force, and the institutional style of draft rulings. Notwithstanding these strengths, several recurring limits emerge. The model's dependence on the factual record and its embedded context, its tendency to generate arguments (even when artificial) for clear cases, the rigidity imposed by a binary template on disputes requiring graduated remedies, and the potential to suggest unrequested sanctions, together delineate both the operational horizon and improvement margins for such systems.

There is no evidence, in our setting, of systematic bias favouring either users or platforms. When the facts genuinely sustain both sides, symmetry is strong; where one side is indefensible, the model struggles to manufacture plausible arguments, as expected from a rational assistant. Some of the limits described point to necessary evolutions: hybrid prompting that integrates controlled context retrieval, the ability to generate more than two remedies or to offer minimalistic rulings in clear-cut cases, and improved filters against procedural overreach.

More broadly, our results support the conclusion that LLM-assisted decision support can play a credible role in the emerging landscape of ODS-based online speech governance. As new applications and institutional designs proliferate under the DSA, the focus will shift from establishing technical feasibility to calibrating, integrating, and governing these rhetorical machines as part of a new procedural ecology [24, 28]. Future work will explore more granular forms of reasoning, adversarial and multi-sided decision models, alignment with evolving human rights doctrine, and the operational integration of AI in the real-world workflow of certified ODS bodies.

References

[1] European Parliament, the Council of the European Union, Regulation (EU) 2022/2065 of 19 october 2022 on a single market for digital services and amending directive 2000/31/EC (Digital Services Act), Official Journal of the European Union, OJ L 277, 27.10.2022, pp. 1–102 (2022).
URL https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

[2] European Commission, Out-of-court dispute settlement bodies under the digital services act (dsa), last update: 10 March 2026 (2026).
URL https://digital-strategy.ec.europa.eu/en/policies/dsa-out-court-dispute-settlement

[3] D. Holznagel, Art. 21 DSA has come to life, Verfassungsblog: On Matters Constitutional (2024). doi:10.59704/8d27bd5f2320bbae.

[4] L. Gradoni, P. Ortolani, Vying for the scales: Content moderation made in europe after one year of dsa, Verfassungsblog: On Matters Constitutional (2025). doi:10.59704/281ae3f68d546081.

[5] J. van Hoboken, J. Quintais, N. Appelman, R. Fahy, I. Buri, M. Straub (Eds.), Putting the Digital Service Act Into Practice: Enforcement, Access to Justice and Global Implications, Verfassungsbooks, 2023. doi:10.17176/20230208-093135-0.

[6] T. Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media, Yale University Press, 2018. doi:0.12987/9780300235029.

[7] K. Klonick, The new governors: The people, rules, and processes governing online speech, Harvard Law Review 131 (6) (2018) 1598–1670.
URL https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf

[8] J. P. Quintais, N. Appelman, R. O. Fathaigh, Using terms and conditions to apply fundamental rights to content moderation, German Law Journal 24 (5) (2023) 881–911. doi:10.1017/glj.2023.53.

[9] L. Gradoni, P. Ortolani, Moderation made in europe: A look into the future of social media content moderation litigation, Verfassungsblog: On Matters Constitutional (2024). doi:10.59704/e2e54bd5b465b8cf.

[10] E. Douek, The Meta Oversight Board and the Empty Promise of Legitimacy, Harvard Journal of Law & Technology 37 (2) (2023) 1–78. doi:10.2139/ssrn.4565180.

[11] L. R. Helfer, M. K. Land, The meta oversight board's human rights future, Cardozo Law Review 44 (6) (2023) 2233–2301.
URL https://scholarship.law.duke.edu/faculty_scholarship/4432/

[12] P. Ortolani, The resolution of content moderation disputes under the digital services act, Giustizia Consensuale 2 (2) (2022) 533–573.
URL https://repository.ubn.ru.nl/bitstream/handle/2066/289314/289314.pdf

[13] L. Gradoni, Rhetorical machines for human rights-based adjudication, Rivista di diritto dei media 9 (Special Issue II) (2025) 258–279.
URL https://www.rivistadidirittodeimedia.it/rivista/rhetorical-machines-for-human-rights-based-adjudication/

[14] R. Alexy, A Theory of Constitutional Rights, Oxford University Press, 2002.
URL https://global.oup.com/academic/product/a-theory-of-constitutional-rights-9780199584239

[15] R. Alexy, Constitutional rights, balancing, and rationality, Ratio Juris 16 (2) (2003) 131–140. doi:https://doi.org/10.1111/1467-9337.00228.

[16] J. Lai, W. Gan, J. Wu, Z. Qi, P. S. Yu, Large language models in law: A survey, AI Open 5 (2024) 181–196. doi:10.1016/J.AIOPEN.2024.09.002.

[17] J. D. Gutiérrez, Critical appraisal of large language models in judicial decision-making, in: R. Paul, E. Carmel, J. Cobbe (Eds.), Handbook on Public Policy and Artificial Intelligence, Edward Elgar, 2024, pp. 323–340. doi:10.4337/9781803922171.00033.

[18] J. Z. Liu, X. Li, How do judges use large language models? Evidence from Shenzhen, Journal of Legal Analysis 16 (1) (2025) 235–262. doi:10.1093/jla/laae009.

[19] L. Gradoni, P. Ortolani, Applicable Law in Out-of-Court Dispute Settlement: Three Vertigos under Article 21 of the DSA (2025).

URL https://orbilu.uni.lu/handle/10993/65952

[20] European Union, Charter of Fundamental Rights of the European Union (2012/c 326/02), Official Journal of the European Union, OJ C 326, 26.10.2012, pp. 391–407 (2012).
URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:C2012/326/02

[21] Council of Europe, Convention for the protection of human rights and fundamental freedoms (1950).
URL https://rm.coe.int/1680a2353d

[22] J. Dik, R. Markovich, Modeling judicial discretion with nuanced permissions, in: J. Savelka, J. Harasta, T. Novotná, J. Mísek (Eds.), Legal Knowledge and Information Systems - JURIX 2024: The Thirty-seventh Annual Conference, Brno, Czech Republic, 11-13 December 2024, Frontiers in Artificial Intelligence and Applications, IOS Press, 2024, pp. 48–59. doi:10.3233/FAIA241233.

[23] D. U. S. de la Osa, N. Remolina, Artificial Intelligence at the Bench: Legal and Ethical Challenges of Informing—or Misinforming—Judicial Decision-Making through Generative AI, Data & Policy 6 (2024) e50. doi:0.1017/dap.2024.53.

[24] V. Fikfak, L. R. Helfer, Automating international human rights adjudication, Michigan Journal of International Law 69 (2025) 1–33. doi:0.36642/mjil.46.1.automating.

[25] M. Mancoridis, B. Weeks, K. Vafa, S. Mullainathan, Potemkin understanding in large language models, in: A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, J. Zhu (Eds.), Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025, Proceedings of Machine Learning Research, PMLR / OpenReview.net, 2025.
URL https://proceedings.mlr.press/v267/mancoridis25a.html

[26] User Rights, Published Decisions, last update: 10 March 2026. Accessed: 2024-03-28 (2026).
URL https://www.user-rights.org/en/decisions

[27] Registry of the European Court of Human Rights, Guide on Article 10 of the European Convention on Human Rights (Freedom of Expression), accessed: 2024-03-28 (2022).
URL https://rm.coe.int/guide-on-article-10-freedom-of-expression-eng/native/1680ad61d6

[28] T. Sourdin, Judges, Technology and Artificial Intelligence: The Artificial Judge, Edward Elgar Publishing, Cheltenham, 2021. doi:10.4337/9781788978262.