Linking Transparency and Accountability: Analysing The Connection Between TikTok's Terms of Service and Moderation Decisions

Leonard Esser **Gerasimos Spanakis**
Department of Advanced Computing Sciences
Maastricht University
{l.eer@student., jerry.spanakis@}maastrichtuniversity.nl

Abstract

The European Commission's Digital Services Act (DSA) mandates that Very Large Online Platforms (VLOPs), like TikTok, provide Statements of Reason (SoRs) to justify their content moderation decisions in an attempt to enhance transparency and accountability for these platforms. However, we can often notice a gap between these automated decisions and the platform's written policies. This leaves users unable to understand the specific rule they have violated. This paper addresses this gap by developing and evaluating a pipeline to link TikTok's SoRs from the DSA transparency database to the most relevant clause from TikTok's policy documents. We test multiple methods to perform the linking task and evaluate performance using a wide range of retrieval methods and metrics.

We develop and deliver a gold-standard dataset where a team of legal research assistants annotated 100 SoRs based on four criteria: clarity, understanding, presence of unclear terms and level of detail, each rated on a 1–4 scale. In addition, a binary rating is assigned for redress clarity. Moreover, annotators determined the best link to the relevant TikTok policy clauses. Results show that both TikTok's SoRs and policy clauses are often extremely broad, granting TikTok more freedom to decide how to apply the clauses, making it even less transparent for users. We also provide a demo that, for each SoR, provides a ranking of the most relevant clauses from TikTok's written policies, a tool that can be useful for users, regulators and researchers to better understand content moderation decisions, assess compliance with transparency requirements, and support further analysis of platform accountability.

1 Introduction

Large online platforms have become a staple part of everyday life for sharing discourse, emotions and social interaction for billions of users. In 2025, it is projected that about 5.24 billion people use social media daily[1] and TikTok alone has 1.12 billion monthly users, spending an average of 95 minutes on the platform[2]. To manage the massive volume of user-generated content, these platforms increasingly rely on automated systems for content moderation (Gillespie, 2018). While this reliance is necessary to counter harmful content, these automated "black box" decisions lead to concerns about fairness, accountability, and transparency (Klonick, 2017).

To address this and enforce greater responsibility, the European Commission introduced the Digital Services Act (DSA)[3]. It is set out to increase the accountability and interpretability for these decisions by making VLOPs publish SoRs that explain why actions like removal or restrictions were taken and what means were used for their detection.

In practice, this link between a specific Statement of Reasons (SoR) and the exact policy clause is rarely clear to users or even legal teams. TikTok's SoRs are often highly templated and vague, which makes it difficult for users and researchers to connect them to the governing rules (Kaushal et al., 2024). This paper addresses the gap in interpretability. Using information retrieval techniques, we develop a pipeline to automatically link TikTok's SoRs from the DSA Transparency Database to the most relevant clauses in its policy documents.

The contributions of this paper are as follows: (a) We create a gold-standard dataset of 100 SoRs (TikTok-100) manually annotated by a team of legal research assistants. For each SoR we evaluate its clarity, understanding, detail level, inclusion of unclear terms as well as the most relevant chunk (out of 124 in total) from the TikTok policy docu-

[1] https://www.demandsage.com/social-media-users/
[2] https://backlinko.com/tiktok-users
[3] https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 382–395

ments, (b) We evaluate traditional sparse retrieval models, like TF-IDF or BM25 and modern dense embedding models, like BERT, OpenAI's embeddings or cross-encoders, and generative models like GPT-4.1, as well as hybrid, fusion and fine-tuning strategies for the task of linking a SoR with platform policy document clauses. (c) We provide a working demo that, for each SoR, orders the most relevant platform policy clauses (for different retrieval models) and incorporates a two-stage fairness assessment pipeline, combining the CLAUDETTE model from (Lippi et al., 2019) with a custom model to flag policy clauses that may potentially be unfair or ambiguous.

2 Background

Content moderation is essential for social media platforms, which act as "new governors" of online speech by setting and enforcing rules (Gillespie, 2018; Klonick, 2017). The DSA tried to make this governance more transparent by demanding that platforms publish SoRs. Under the EU Digital Services Act (DSA), providers of online platforms must issue statements of reasons when moderating content (Art. 17), including the contractual or legal ground relied upon, whether automated tools were used, and available redress mechanisms. Furthermore, platforms must submit these SoRs to a publicly available database (DSA Transparency Database (Art. 24(5)), which exposes large-scale, near-real-time moderation rationales.[4] However, initial analyses of the DSA database reveal that platforms like TikTok often use repetitive, vague statements that undermine fairness and accountability (West, 2018; Shahi et al., 2025). For instance, TikTok frequently cites generic "Community Guidelines" violations and relies heavily on automated tools for over 95% of decisions, leading to standardised explanations lacking specific detail.[5] Early analyses of the DSA database already confirm this trend and show the differences within moderation practices. A study of over 156 million SoRs by (Drolsbach and Pröllochs, 2024) found that TikTok is by far the most active platform and performed over 350 times more moderation actions per user than X (Twitter), with the majority of decisions being automated.

The DSA transparency database was designed to counter exactly this kind of behaviour from platforms. These are more and more shifting their practices from removing content outright to new methods that hinder visibility for users. In fact, it can be argued that the DSA's requirement to provide a SoR for every moderation action could function as a "prohibition on shadow banning". (Leerssen, 2023) As we can see later, however, guidelines like these actually lead to the problem that is already discussed in the paper itself, in that platforms use visibility reduction for ambiguous content and, in that way, create a situation where the most sensitive cases are governed by the least transparent means.

Prior work has also highlighted discrepancies between platforms' stated policies and their practices. This tendency is reflected in platforms' official transparency reports, too. An analysis of these reports by Urman and Makhortykh (2023) found that companies are much more willing to talk about government removal requests than their own moderation decisions, which remain "largely obscure". The study argues that this can be seen as a form of "transparency-washing", which looks like they are open about their rule enforcement, while in reality, it stays mostly obscure to users. TikTok has faced criticism for unclear moderation practices, such as allegedly suppressing content from creators deemed "ugly, poor or disabled" (Zeng and Kaye, 2022). And while TikTok seems to provide more detail than some rival companies like Facebook (or Meta), which reportedly cited a generic *"violation of our terms"* for almost 100% of its removals, TikTok still uses a "similar albeit shorter statement" across cases (Kaushal et al., 2024). This means that users usually get a standardised repetitive block of text, instead of an actually helpful reason.

The discrepancy is also found in audits of the DSA database itself. For example, an analysis by (Trujillo et al., 2025) of over 350 million SoRs found "striking inconsistencies" between the data that platforms submitted to the database and the information they stated in their own reports. The most significant contradictions were in the use of automation, where X (formerly Twitter) reported using no automation at all in the database despite saying otherwise in their reports.

NLP techniques have been used to improve the interpretability of legal documents, addressing a core reason that users perceive something as unfair, which is the lack of clear, consistent linking between a moderation decision and the platform's own rules. For instance, a study of YouTube cre-

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065

[5] https://newsroom.tiktok.com/en-eu/dsa-third-transparency-report

ators by (Ma and Kou, 2022) found that the perception of users and what they see as unfair relies heavily on the consistency of moderation and the equality when compared to other creators. If they feel like another user's content is not removed, even though they made the same content as them, they often deem the process unfair and arbitrary.

The CLAUDETTE system uses machine learning to automatically detect potentially unfair clauses in ToS documents (Lippi et al., 2019). More recently, Aspromonte et al. (2024) used a multi-agent system with LLMs to link SoRs to ToS clauses. This approach can be computationally expensive, however, and can lead to error propagation. Our work builds on these findings by providing a broader comparative analysis of a number of different retrieval methods, including sparse, dense, and hybrid models, and integrating a fairness assessment pipeline specifically for the TikTok statements and clauses.

3 Data

Our work relies on three primary data sources: TikTok's moderation decisions (SoRs), its policy documents, and a manually annotated gold-standard dataset that we use for the evaluation.

3.1 DSA Transparency Database

We collected approximately 1.2 billion SoRs submitted by TikTok to the DSA Transparency Database.[6]Each SoR contains up to 37 fields, but our analysis focuses on the **"incompatible_content_explanation"** field, which contains TikTok's justification for the moderation action. Our analysis confirmed that the explanations are highly repetitive. The single most common explanation, related to harassment and trolling, accounts for over 36% of all entries, and the top 10 unique explanations cover over 85% of the dataset. This really shows the templated nature of TikTok's transparency reports. An example of a SoR can be found in Appendix A.

3.2 TikTok Policy Documents

To create a corpus that is as complete as possible with TikTok's rules, we combined five key documents:

1. **Terms of Service (ToS):** The core legal contract for the EEA/UK/CH. Other regional variants, like the US one, differ in wording and scope.[7]
2. **Community Guidelines:** Concrete "dos and don'ts" for creators specifically.[8]
3. **TikTok Ad Policies:** Specific rules for features like Rewards and Music.[9]
4. **Brand Guidelines:** Rules for sponsored or branded content.[10]
5. **Commercial Terms:** Rules for advertisers using TikTok's ad platform and businesses.[11]

In order to be able to link to specific segments of these legal documents later on, we segmented the combined texts into logical "chunks" representing individual clauses or paragraphs. We first experimented with rule-based methods, like splitting by markdown headings or newlines, but we found that these methods produced inconsistent and often logically unclear chunks.

Clause-level segmentation is also an option for some of the documents, like the ToS, but after testing it (also) led to largely inconsistent results, as some documents are not clearly segmented by clauses. Furthermore, some clauses grouped together by TikTok in those documents were very long, multi-topic, or structured as open-ended bullet lists with cross-references. This would, even if a linkage succeeded, lead to an unclear result for the user. For that reason, we opted to use OpenAI's GPT-4.1 in combination with the use of TikTok's own headline structure, where available, to perform the chunking into logical chunks that keep statements about one topic together while avoiding overly broad segmentations. The AI prompt can be seen in the appendix. This yielded 124 distinct chunks. When creating these chunks, it already became clear that some of the chunks consist of very broad "catch-all" phrases. For example, chunks that provide a massive list of things that you are not allowed to advertise. For users, it is then hard to grasp which of these things they violated.

3.3 TikTok-100: Gold-standard dataset

To create a gold standard for evaluation, we randomly selected 100 unique SoRs from our dataset. Each SoR was independently annotated by two (out

[6] https://transparency.dsa.ec.europa.eu/

[7] https://www.tiktok.com/legal/page/us/terms-of-service/en

[8] https://www.tiktok.com/community-guidelines/en

[9] https://ads.tiktok.com/help/article/tiktok-advertising-policies

[10] https://tiktokbrandhub.com/legal

[11] https://ads.tiktok.com/i18n/official/policy/commercial-terms-of-service

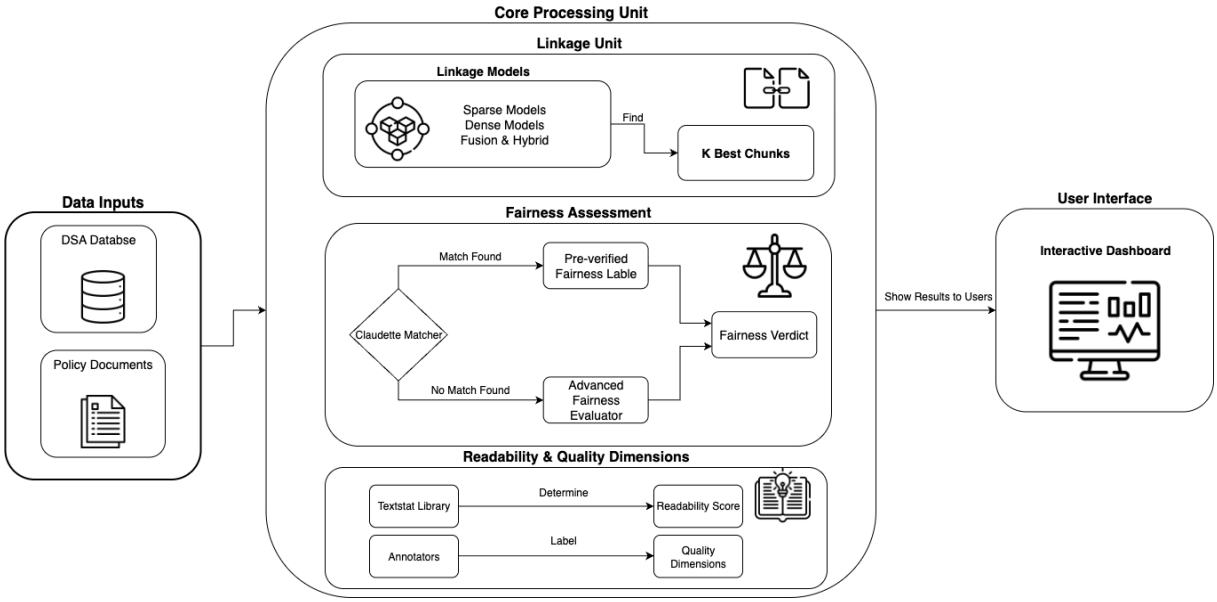

Figure 1: Overall Structure of the Application

of four in total) legal research assistants. For each SoR, annotators were tasked with:

1. **Selecting the best-matching policy chunk(s)** from our corpus of 124 chunks.
2. **Rating the SoR on four clarity dimensions** (Clarity, Understanding, Unclear Terms, Detail Level) on a 1-4 scale, plus a binary rating for Redress Clarity.

This process yielded 200 total annotations, forming the basis for evaluating our automated models and analysing the ambiguity of the linkage task itself.

Agreement between legal research assistants was limited, with Cohen's kappa averaging 0.243, indicating slight reliability despite a raw agreement rate of 68.22%. This suggests that the task is extremely ambiguous, often because multiple policy chunks are plausible matches for a single vague SoR and because TikTok uses so many "catch-all" clauses, which explain the high difference between the relatively high overall agreement and the low Cohen's kappa scores. The agreement for binarised clarity dimensions was higher, like 82% for Understanding, but nearly all annotations (98%) agreed that SoRs fail to provide clear information on redress options. Detailed results for the annotation definitions, as well as the experiments, can be found in Appendix C. We also release the full data and all annotations[12].

[12] https://github.com/Leonard-git-things/Transparency-EMNLP

4 Methodology

Our application pipeline can be seen in Figure 1. In this section, we will describe the core processing unit, comprising the linkage unit (retrieval models used for linking SoRs with ToS), the fairness assessment and the readability. Details for the implementation of the application (demo) and screenshots can be found in Appendix G.

4.1 Linkage unit: Linking SoRs and ToS

We formulate the problem of linking SoRs to ToS as an information retrieval problem, where a SoR's explanation is the query and the 124 policy chunks form the retrieval pool. We compare several retrieval models.

Sparse Retrieval (Lexical): These models rely on keyword matching. We used *TF-IDF* and *BM25* as strong and, importantly, also transparent baselines. BM25 enhances the capabilities of TF-IDF by using parameters to account for the term frequency and address document length normalisation. This often improves performance on short queries.

Dense Retrieval (Semantic): These models capture semantic meaning by encoding text into dense vector representations. We evaluated a number of models, including general-purpose *BERT* (Devlin et al., 2019), *DPR* (Karpukhin et al., 2020), the domain-specific *LegalBERT* (Chalkidis et al., 2020), and OpenAI's powerful *"text-embedding-3-large"* model. We also tested a *Cross-Encoder* model, based on the BERT model, which processes the SoR and chunk pair at the same time in an

attempt to gain a deeper context understanding.

Generative Models: We prompted Large Language Models (*GPT-4o*, *GPT-4.1* and *GPT-o4-mini*) to perform the linkage in a zero-shot, forced-choice setting, where the model was asked to return the ID of the single best-matching chunk from the provided corpus. (Aspromonte et al., 2024)

Hybrid Strategies: Following (Louis et al., 2025), we tested hybrid and fusion techniques to assess the performance gains from combining multiple models, particularly sparse and dense ones. For our experiments, we used TF-IDF and BM25 as the sparse models and the embedding model by OpenAI as the dense model, as they had the best individual performance.

- **Hybrid Retrieval (Early Fusion):** For this, we compute a unified score via linear interpolation: $S_{hybrid} = \alpha \cdot S_{sparse} + (1-\alpha) \cdot S_{dense}$, where α indicates the influence of the sparse and dense models (Louis et al., 2025). The bigger α is, the greater is the sparse model's influence.
- **Late Fusion:** To achieve this, we combined the ranked lists from multiple individual models using methods like *Reciprocal Rank Fusion (RRF)*, *Majority Voting*, *Score Aggregation*, *Score Interpolation*, *Ensemble Fusion*. A description of these methods can be found in the appendix under the section D.

4.2 Fairness and Clarity Assessment

Fairness: We developed a two-stage pipeline to flag potentially unfair clauses. First, a *CLAUDETTE-based matcher* finds chunks from TikTok's ToS that have been annotated as unfair by the CLAUDETTE model (Lippi et al., 2019). They label unfair categories like Unilateral Termination and Limitation of Liability. For clauses not found in CLAUDETTE, a custom *Advanced Fairness Evaluator (AFE)* applies a rule-based system using weighted regular expressions to detect patterns that often appear in clauses that are potentially unfair. Examples would be "at our sole discretion" or "without prior notice". More details can be found in Appendix E.

Quality Dimensions: After binarising the ratings from our annotations in the gold-standard dataset, we trained logistic regression classifiers to predict the values of a given SoR across the four dimensions: Clarity, Understanding, Unclear Terms, and Detail Level. Notably, we leave out the redress dimension here, as there were almost no positive labels in all of the 100 SoRs. We deliberately use simple, interpretable classifiers because the dataset is small and heavily imbalanced (e.g. near-zero positives for redress). The models also work towards our goal of providing a more transparent baseline and not introducing more uncertainty.

4.2.1 Readability

To provide even more help to users for understanding a given policy chunk, our interactive dashboard includes a readability feature, which uses the *textstat* library in Python to generate a readability score for the chunk. Specifically, *textstat.text_standard*, which uses a combination of several readability tests like the *Flesch-Kincaid Grade Level* (Solnyshkina et al., 2017) or the *SMOG index* (Mc Laughlin, 1969), and returns an estimated school grade level required to understand the text. This helps the users to quickly see its complexity. This component was purely for the demo and, therefore, not included in our formal evaluation.

5 Results and Analysis

We evaluated our models against the TikTok-100 dataset (§3.3). We use standard metrics for effectiveness, namely mean reciprocal rank (MRR) as a rank-aware metric and recall at various thresholds (R@k), which ignores rank but can be particularly useful for assessing performance in an ambiguous task like ours. More detailed descriptions of these methods can be found in Appendix F.

5.1 Retrieval Performance

We experiment with zero-shot retrieval for individual and hybrid/fusion models (§5.1.1 and with fine-tuning on our dataset (§5.1.2), the latter being more of a proof-of-concept due to the small size of our dataset.

5.1.1 Zero-shot retrieval results

As we can see from the first part of Table 1, the general-purpose OpenAI embedding model achieves the highest performance with an MRR of 0.691. Notably, the sparse model BM25 is also highly competitive, while the dense models surprisingly seem to lag behind. Generative models also exhibited relatively weak performance.

In Figure 2, we show the recall performance of individual models for different k values. Most models exhibit a noticeable jump after the first few

Model	MRR	R@1	R@5	R@20
Individual Models				
OpenAI Embedding	0.6911	0.5556	0.8778	0.9556
BM25	0.6787	0.5778	0.8222	0.9222
TFIDF	0.6504	0.5000	0.8333	0.9889
GPT-4o	0.4000	0.4000	0.4000	0.4000
DPR	0.3428	0.1667	0.5444	0.9000
CrossEncoder	0.2536	0.0444	0.4889	0.8222
GPT-4.1	0.2500	0.2500	0.2500	0.2500
BERT	0.2441	0.1556	0.3333	0.7333
o4-mini	0.1000	0.1000	0.1000	0.1000
LegalBERT	0.0861	0.0667	0.1000	0.1556
Hybrid/Fusion Models				
Hybrid BM25 ($\alpha = 0.2$)	0.7841	0.7111	0.8889	0.8889
Hybrid BM25 ($\alpha = 0.3$)	0.7606	0.6889	0.8778	0.8778
Hybrid TFIDF ($\alpha = 0.3$)	0.7587	0.6667	0.9000	0.9000
Hybrid TFIDF ($\alpha = 0.4$)	0.7500	0.6444	0.9000	0.9000
Hybrid TFIDF ($\alpha = 0.7$)	0.7226	0.6000	0.9111	0.9111
Score Interp. (BM25+OpenAI)	0.7146	0.6111	0.8667	0.8667
Score Interp. (TFIDF+OpenAI)	0.7128	0.6000	0.8889	0.8889
Ensemble Fusion	0.7020	0.5556	0.8889	0.9778
Majority Voting	0.6903	0.5444	0.8778	0.9778
Hybrid BM25 ($\alpha = 0.7$)	0.6730	0.5556	0.8556	0.8556
Score Aggregation (Avg)	0.6562	0.5222	0.8556	0.9667
Rank Fusion (RRF)	0.5025	0.2778	0.8222	0.9667

Table 1: Zero-shot results of Individual and Fusion Models, each section ranked by MRR.

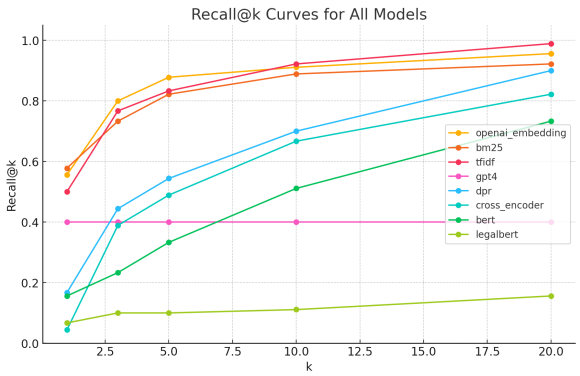

Figure 2: Recall@k Curves for All Individual Models

k values, reflecting cases of ambiguity: when annotators linked one chunk but a model ranked a different chunk higher, both may actually be correct, yet only one was chosen as the "best-fitting" reference. These apparent errors occur due to such ambiguity and diminish as k increases, indicating that the method is generally able to capture relevant alternatives when allowed to consider more candidates.

The lower part of Table 1 shows that hybrid strategies boost performance, since in most cases these methods outperform the best individual models. The most effective method was the hybrid combining BM25 and OpenAI embeddings, which is consistently at the top of the leaderboards. It achieved an MRR of 0.784. After a thorough analysis that compared the results of using different α values, we find that $\alpha = 0.2$ is the best balance between sparse and dense models. A similar hybrid model used TF-IDF and also performed exceptionally well with an MRR of 0.759. These results strongly suggest that the ideal approach is neither purely lexical nor semantic but requires a blend between those two.

The late fusion methods notably also performed better than almost all the individual models in themselves, but were still worse than the strongest individual models. The strongest model here was *majority voting* with an MRR of 0.688. This shows that even when we don't do an early fusion, a hybrid approach, combining several models, still leads to more robust and accurate linking, all in all.

Further analysis of only looking at partial fusions revealed that more is not always better. A targeted fusion that looks at only a few models instead of always combining all of them achieved a higher MRR, suggesting that careful model selection is often more effective than quantity. Results can be seen in Table 2. In this Table *Cross-Domain Pair* stands for a **late fusion** of a sparse model (BM25) and one dense model (OpenAI embedding). *Sparse + Dense* combines two sparse and one dense model. *Balanced mix* uses a set of four models with two sparse and two dense.

Table 2: Performance of targeted partial fusion methods compared to comprehensive fusions and individual models.

Method	Type	Count	MRR	R@5
Rank Fusion (RRF) - Cross-Domain Pair	Fusion	2	0.770	0.922
Score Aggregation - Cross-Domain Pair	Fusion	2	0.747	0.878
Majority Voting - Sparse + Dense	Fusion	3	0.744	0.878
Majority Voting - All Models	Fusion	6	0.700	0.878
Majority Voting - Balanced Mix	Fusion	4	0.694	0.867
OpenAI (Best Single Model)	Individual	1	0.691	0.878
BM25	Individual	1	0.679	0.822
Rank Fusion (RRF) - All Models	Fusion	6	0.662	0.833

5.1.2 Retrieval Fine-Tuning Results

We are also looking into the impact of supervised fine-tuning on two of the dense models, namely BERT and DPR. For this purpose, we split the dataset of the 100 SoRs into 80 used for training and 20 used for testing. We are aware that the small dataset does not capture the complexity of the task; however, this experiment serves as a proof of concept for future applications. As we can see in Figure 3, the results peaked after only a few epochs of training. We believe that the results follow the trend observed previously, i.e. the massive use

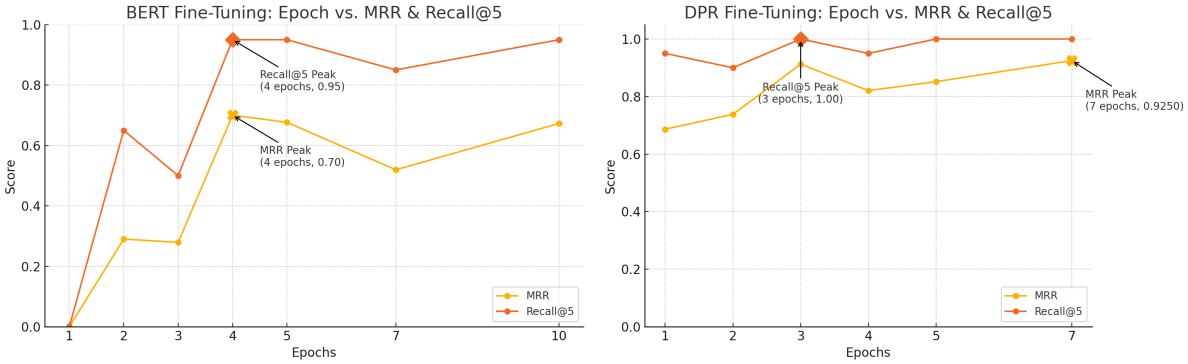

Figure 3: Recall and MRR curves of fine-tuned retrieval models. Left: BERT. Right: DPR.

of "catch-all" clauses that the models then learn to link to when in doubt. Importantly, this does not indicate the failure of the model, but rather a shortcoming of the transparency that TikTok should give its users within the SoRs, but does not. As we saw above, even for the human legal research assistance annotator team, it was difficult to always find one definite chunk to link to, as the task was so highly ambiguous.

5.2 SoR Quality Dimensions Classification

Besides retrieval, we classified the four quality dimensions into "low" (average rating up to 2.5) and "high" (average rating above 2.5). After filtering out entries that contained errors or were mistakenly not labelled by the annotators, this left us with a total of 94 samples. This means a training set of 75 samples and a test set of 19.

As expected, the dataset is highly imbalanced. For example, *Clarity* and *Unclear Terms* had "high" classifications in 96% of cases, while *Detail Level* was the most balanced with 65% "high" classifications. We trained and evaluated three different logistic regression classifiers (Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM)) using the SMOTE oversampling technique (Chawla et al., 2002).

Results for both training and test sets can be seen in Table 3. As expected, we see that for the more balanced dimensions like *Detail Level* and *Unclear Terms*, performance on the test set drops, a sign of overfitting. On the other hand, for the highly imbalanced dimensions *Clarity* and *Understanding*, the models seem to achieve perfect or almost perfect F1-scores on the test set, indicating (also after some manual inspection) that the model learns trivial solutions. Due to the massive class imbalance and the small size, the test set was not diverse enough to include examples that consequently proved this simple rule wrong. Both findings show that the classifiers show promise, but a larger annotated dataset would be needed to build more robust models.

Classifier	Algorithm	Test_AUC	Test_F1	Train_AUC	Train_F1
Clarity	LR	0.947	0.973	0.996	0.987
Clarity	RF	0.947	0.947	1.000	1.000
Clarity	SVM	0.947	0.973	1.000	0.993
Detail Level	LR	0.637	0.828	0.942	0.914
Detail Level	RF	0.527	0.692	0.998	0.990
Detail Level	SVM	0.560	0.828	0.967	0.942
Unclear Terms	LR	0.789	0.889	0.994	0.993
Unclear Terms	RF	0.816	0.889	0.998	0.993
Unclear Terms	SVM	0.684	0.857	0.994	0.993
Understanding	LR	1.000	1.000	0.985	0.971
Understanding	RF	1.000	1.000	0.999	0.993
Understanding	SVM	1.000	1.000	0.992	0.985

Table 3: Model performance of classifiers on test vs training set

5.3 Error Analysis

To better understand the model performance outside of the standard metrics, we conducted an error analysis. Our review showed that many of the apparent "failures" were not incorrect linkages but rather selections of semantically similar clauses.

From a **quantitative analysis**, we looked into the failure overlap, in order to see whether there is any pattern. As we can see in Figure 4, models with similar architectures tend to struggle on the same types of SoRs. For example, BERT-based models (like BERT, LegalBERT, and DPR) show a high error overlap, and so do the two sparse models, TF-IDF and BM25. This shows that errors are systematic and tied to specific model limitations rather than being random. We also found that model performance was largely unaffected by the amount of human annotation agreement on a single best chunk. Models, therefore, are robust to the inherent ambiguity of the task.

Looking into **qualitative insights**, we reviewed the most challenging SoRs (i.e. those that most

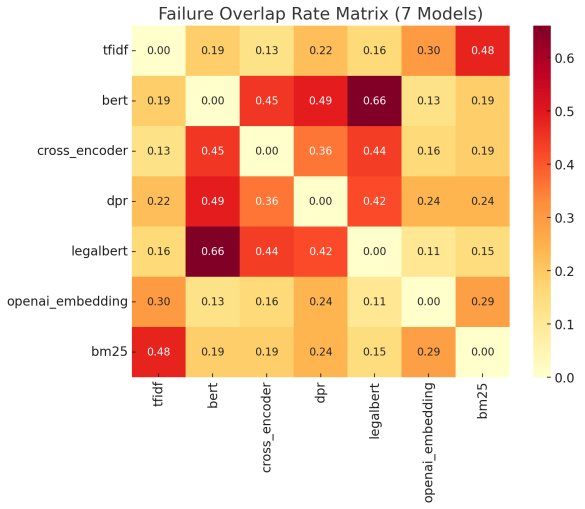

Figure 4: Retrieval Model Failure Overlap

models failed). This process revealed the ambiguity of the task, where in a lot of instances, cases that were marked as "wrong" might not have been completely "wrong" at all. As mentioned in §3.2, we combined 5 different policy documents, which led to some chunks being semantically similar, depending on which document they are referenced. An example of this is the reference to the minimum age of users for TikTok. The stated reason by TikTok in its SoR is: *"You must be 13 years and older to have a TikTok account, and be 18 years and older to go LIVE. There are additional age limitations based on local law in some regions. We are deeply committed to ensuring that TikTok is a safe and positive experience for people under the age of 18. If we learn someone is below the minimum age on TikTok, we will ban that account. If we learn someone is below the minimum age to go LIVE, we will ban their LIVE Access."*

We can see that models have several options to which chunk to point to for this, like *"[...] Users must be at least 13 years of age to have an account. However, additional age limitations may apply based on local laws in some markets [...]"* from *TikTok's ad policies* or *"Minimum age: You must be 13 or older to use the Platform. Accounts for users found to be underage will be terminated. Appeals available for mistaken termination."* from *TikTok Terms of Service*. The SoRs do not provide access to the actual content; therefore, it is impossible to know which policy applies here. That also aligns with findings from Figure 2: For moderately big numbers of k, we almost guarantee that the model picks the correct chunk, while the ranking might differ. However, this makes the linking/ranking tool/application useful for users and/or regulators so they can inspect the final result.

Another example where the models failed was the statement *"Many people around the world find entertainment through games of chance. While TikTok is an entertainment platform, we recognise that risking money in a game or a bet may lead to potential harm for some people, including serious financial loss or addiction. We do not allow the promotion of gambling services. Users and 3rd-parties can report policy violations to us. We have detected this policy violation based on a report that the content violated our Community Guidelines."*. Same as before, there are many chunks that reference gambling or games with chance in some way, but it is hard for models to find out which one is the best-fitting one without knowing what the removed content was. The annotators did not agree either, but both found chunks that make sense to include: For example, one annotator referenced *"[...] We prioritise audience safety by regulating gambling and related activities. [...]"* and the other referenced a chunk that lists all the things that are forbidden when making branded content, and gambling was one item amongst that.

Overall, our error analysis reveals a critical insight: the models' "failures" are a clear sign of the lack of clarity in TikTok's policies and statements. The ambiguity is not necessarily a failure of the model but a result of the platform's failure to provide clarity to the users. Because of this gap in unambiguous wording, users and automated systems alike face problems in interpreting and applying these rules consistently.

6 Conclusion

This paper presented a comprehensive pipeline for linking TikTok's moderation decisions to its policies. This is an important step towards enforcing the transparency that the Digital Services Act originally mandated. Our evaluation across a wide range of retrieval models showed several key insights. First, hybrid retrieval strategies that fuse sparse and dense methods are overall the most effective, outperforming any individual model. Second, general-purpose models provide stronger out-of-the-box performance than domain-specific ones like LegalBERT for this task. Third, supervised fine-tuning provides significant performance gains. For further research, it would be interesting to ex-

plore what effects it would have to create a larger, legally sound annotated dataset to avoid overfitting and get more meaningful insights. The same applies to assessing the fairness and clarity of the clauses.

Overall, our analysis confirms that TikTok -like many other platforms- relies on vague and repetitive explanations as well as overly broad "catch-all" clauses that obscure the real reasoning behind moderation decisions. This then creates a gap regarding accountability. By automatically linking the moderation practice to the policies and flagging potentially unfair terms, our work provides a methodology and a practical toolkit for regulators, researchers, and users to better analyse and understand the moderation systems of online platforms.

Limitations

We identify the following limitations. First, our analysis is confined to English-language documents, which may introduce a bias towards moderation patterns in English-speaking regions. The European Union (where DSA applies) is highly multilingual, and only looking at statements in English might lead to the loss of some insightful information. Similarly, the legal texts analysed were EEA/UK/CH facing and might differ from the US ones. Second, our gold-standard dataset, while expertly curated, is small, consisting only of around 100 SoRs, which limits the statistical power of our evaluations, especially for the fine-tuning experiments and the clarity classifications. Third, we only look at TikTok as a platform, and the generalisation to other platforms is untested. Finally, our fairness assessment is an automated indicator based on textual patterns and is not a definitive legal judgment. It might give users a hint about what clauses might be worth appealing to and which ones can be considered fair but should not act as legal ground..

References

Marco Aspromonte and 1 others. 2024. LLMs to the Rescue: Explaining DSA Statements of Reason with Platform's Terms of Services. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 205–215.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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 Web Conference 2024*, pages 939–942.

Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Rishabh Kaushal and 1 others. 2024. Automated transparency: A legal and empirical analysis of the Digital Services Act Transparency Database. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1121–1132.

Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598.

Paddy Leerssen. 2023. An end to shadow banning? transparency rights in the digital services act between content moderation and curation. *Computer Law & Security Review*, 48:105790.

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.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2025. Know when to fuse: Investigating non-English hybrid retrieval in the legal domain. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4293–4312, Abu Dhabi, UAE. Association for Computational Linguistics.

Renkai Ma and Yubo Kou. 2022. "I'm not sure what difference is between their content and mine, other

than the person itself" A Study of Fairness Perception of Content Moderation on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Gautam Kishore Shahi and 1 others. 2025. A Year of the DSA Transparency Database: What it (Does Not) Reveal About Platform Moderation During the 2024 European Parliament Election. *arXiv preprint arXiv:2504.06976*.

Marina Solnyshkina, Radif Zamaletdinov, Ludmila Gorodetskaya, and Azat Gabitov. 2017. Evaluating text complexity and Flesch-Kincaid grade level. *Journal of social studies education research*, 8(3):238–248.

Amaury Trujillo, Tiziano Fagni, and Stefano Cresci. 2025. The DSA Transparency Database: Auditing self-reported moderation actions by social media. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–28.

Aleksandra Urman and Mykola Makhortykh. 2023. How transparent are transparency reports? comparative analysis of transparency reporting across online platforms. *Telecommunications policy*, 47(3):102477.

Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383.

Jing Zeng and D Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, 14(1):79–95.

A Statement of Reason Example

Statement of reason details: 66a1d177-03cb-41ea-bf56-07cbb047afba

Platform name	TikTok
Received	2025-08-09 23:59:33 UTC
Visibility restriction	Removal of content
Facts and circumstances relied on in taking the decision	The decision was taken pursuant to own-initiative investigations.
Decision Ground	Content incompatible with terms and conditions
Reference to contractual ground	Youth Exploitation and Abuse
Explanation of why the content is considered as incompatible on that ground	Allowing young people to explore and learn safely during their unique phase of development is our priority. We do not allow youth exploitation and abuse, including child sexual abuse material (CSAM), nudity, grooming, sextortion, solicitation, pedophilia, and physical or psychological abuse of young people. This includes content that is real, fictional, digitally created, and shown in fine art or objects.We proactively enforce our Community Guidelines through a mix of technology and human moderation. We have detected this policy violation using automated measures. We have used automated measures in making this decision.
Is the content considered as illegal?	N/A
Territorial scope of the decision	Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden
Content Type	Other
Content Type Other	Photo Comment
When the content was posted or uploaded	2025-08-09
Category	Protection of minors
Information source	Own voluntary initiative
Was the content detected/identified using automated means?	Yes
Was the decision taken using other automated means?	Fully automated
Application date of the decision	2025-08-09

Figure 5: An Example SoR from the DSA

B Chunking Prompt

The AI prompt for chunking the documents into logical chunks:

You are an AI that can logically chunk long text into meaningful sections. Given the following Terms of Service/legal content documents, break them into logical chunks. For each chunk, output an ID (starting at 1), a Title that summarizes the chunk, the chunked important data (key details), and a very, very short description (a few words). Output the result strictly as CSV with the columns: ID; Title; Chunk; Description. The delimiter of the CSV should be a semicolon (;). Do not include any extra commentary or formatting.

C Annotation Experiment

C.1 Questions and Clarification

1. **Clarity Rating**
   - *Scale*: 1-4 (1 = very unclear, 4 = very clear)
   - *Question*: "How clear is this explanation?"
   - *Meaning*: language, structure, flow
2. **Understanding Rating**
   - *Scale*: 1-4 (1 = very difficult, 4 = very easy)

- *Meaning*: content, semantics, legal requirements, understandability
- *Question*: “Is the rule that is the basis of the decision explained well?”

3. **Redress Clarity**
   - *Options*: Yes = 1/No = 0/Unsure
   - *Question*: "Is the possibility of redress clearly given?"
   - *Meaning*: Whether the statement of reason includes information on the possibility to redress. E.g. does it say that within a few weeks you have to email them for them to reconsider the decision

4. **Unclear Terms**
   - *Scale*: 1-4 (1 = heavy jargon/unclear terms, 4 = no jargon/very clear language)
   - *Question*: "How much unclear jargon or technical terms are used?"
   - *Meaning*: unclear specific words, technical terminology etc.

5. **Detail Level**
   - *Scale*: 1-4 (1= very unclear why it breached the rule; 4 = very clear why it breached the rule)
   - *Question*: "Is the explanation detailed enough?"
   - *Question*: "How easy is it to understand why the content was removed?"
   - *Meaning*: whether TikTok included the explanation of why the act

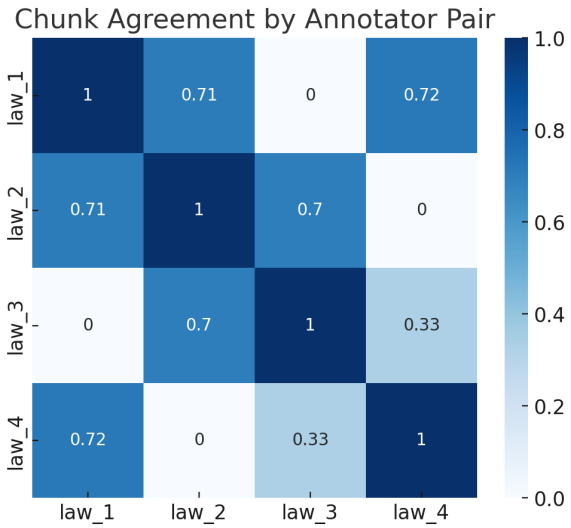

Figure 6: Absolute Agreement per Chunk

Table 4: Summary of Ratings and Agreement Levels

Dimension	Mean Score	Standard Deviation	Agreement (%)
Clarity	3.40	0.84	42.00
Understanding	3.08	0.97	43.00
Redress Clarity	0.01	0.10	98.00
Unclear Terms	3.08	0.86	19.00
Detail Level	2.20	0.84	35.00

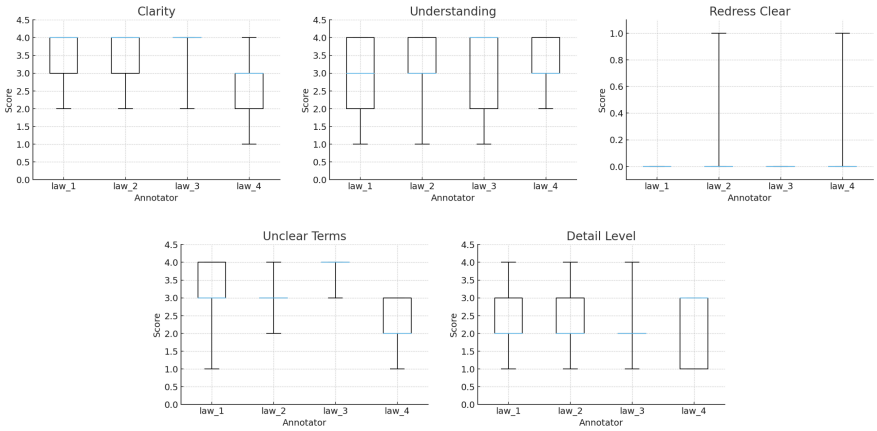

Figure 7: Annotator Bias for Clarity Dimensions

D Fusion Strategies

We assessed the following late fusion methods:

- **Majority Voting:** Ranks the best chunk based on how frequently they appear within the top-ranked results across multiple models.
- **Score Aggregation:** Averages the normalised relevance or retrieval scores from different models to produce one single combined metric.
- **Reciprocal Rank Fusion:** Combines rankings from multiple models by assigning weights inversely proportional to their ranks. This then favours the chunks that are consistently ranked high by a number of different methods.
- **Ensemble Fusion:** A fusion of fusion methods. This method combines the results from voting, scoring and ranking methods and aggregates different fusion strategies into one retrieval ranking.

E Complete Formula for Advanced Fairness Evaluator

$$B = \sum_{i=1}^{n} \big(w_i\, c_i + b_i\big),$$

$$D = \max\big(0,\, 1 - 0.1\,(n-1)\big),$$

$$\delta_s = \begin{cases} +0.10, & \text{if source contains "terms of service" or "commercial terms"} \\ +0.05, & \text{if source contains "community guidelines"} \\ -0.05, & \text{if source contains "advertising policies"} \\ 0, & \text{otherwise} \end{cases}$$

$$\delta_t = \begin{cases} +0.10, & \text{if title contains a warning keyword} \\ 0, & \text{otherwise} \end{cases}$$

$$M = 1 + \delta_s + \delta_t,$$

$$L = \begin{cases} 0.95, & |\text{text}| > 1000, \\ 1.05, & |\text{text}| < 200, \\ 1.00, & \text{otherwise} \end{cases}$$

$$\text{final\_confidence} = \min\big(B \times D \times M \times L,\, 1.00\big).$$

Where:

n the number of matched indicators;

w_i weight of indicator i (e.g. 0.5–0.9);

c_i category weight for indicator i (e.g. 0.5–1.0);

b_i confidence boost for indicator i (e.g. 0.0–0.2);

F Evaluation Metrics

We used the following common information-retrieval and NLP metrics that are described here for completion:

- **Mean Reciprocal Rank (MRR):** Measures the average position of the first correct match out of the top-k given options by the model. An MRR closer to 1 indicates superior performance.
- **Recall@k:** Indicates the percentage of times the correct chunk was included within the top-k predictions made by the model.

G Interactive Transparency Dashboard

To demonstrate the practical applications of our research, we developed a proof-of-concept "transparency dashboard". The dashboard is a web interface intended for users or content creators. The goal is to try to make these people better understand the content moderation decisions of TikTok. It is implemented as a lightweight Flask application and is container-ready via a Dockerfile.

User Workflow and Features

The user is able to see mainly three things in the dashboard:

SoR Lookup: A user can enter a UUID or alternatively the text of any SoR from the DSA database and select a model to perform the linkage and optionally also the number of matches the user wants to look at. The dashboard fetches the result(s) from the model, as well as the entry from the DSA database and displays the result to the user. This can be seen in figure 8.

Clause Linkage and Fairness Verdict: The dashboard then also displays the most relevant policy chunks chosen by the model that correspond to the SoR's explanation. The result is shown with the similarity score, the source document, like the ToS, and a fairness verdict, as discussed in section 4.2. This can be seen in Figure 9.

Model Comparison: Lastly, the dashboard also gives the user the possibility to compare the outputs of a number of linkage models at the same time. The results are presented in a summary table and as an agreement matrix. The matrix shows the user the agreement that different models have on the same statements, which provides the user with even more transparency into the possibilities of the ranking. This can be seen in Figure 10.

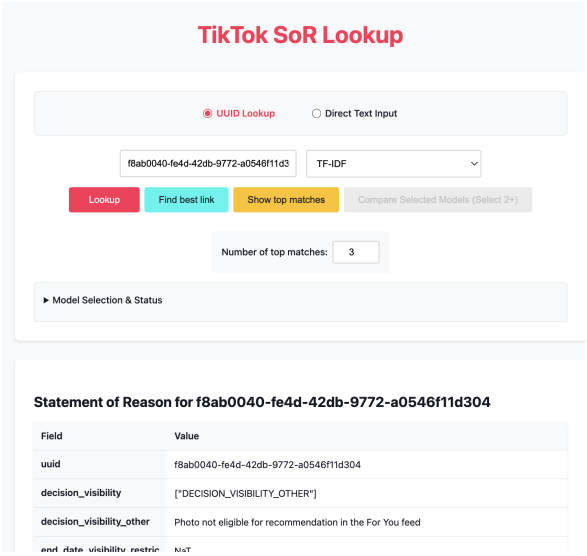

Figure 8: Interface Landing Page

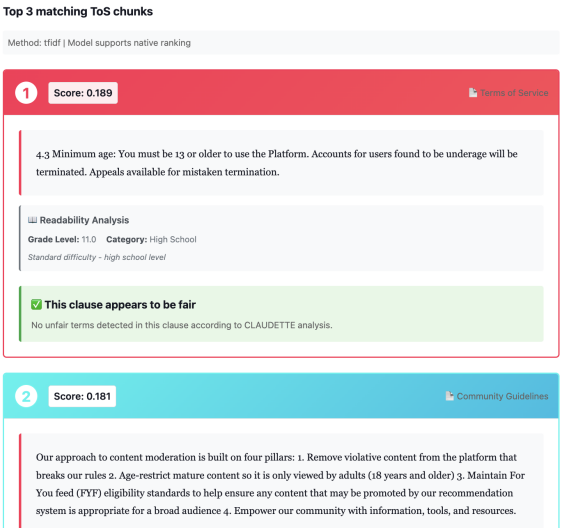

Figure 9: The interface representation of ranking and a potentially fair clause

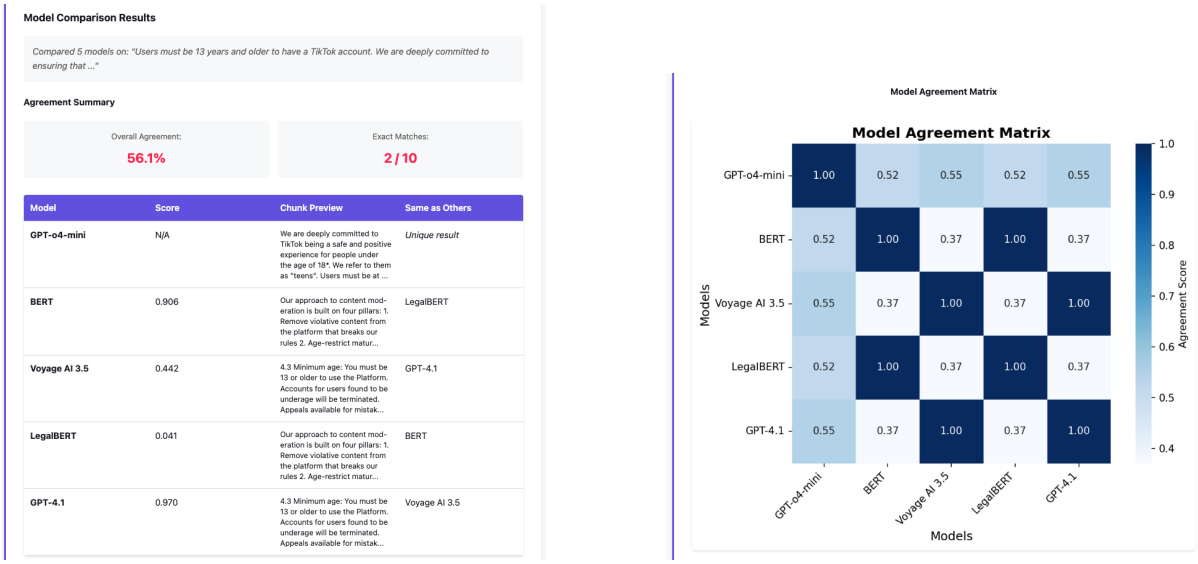

Figure 10: Comparison of model agreement: (a) table view and (b) matrix view.