Evaluating LLM-Generated Legal Explanations for Regulatory Compliance in Social Media Influencer Marketing

Haoyang Gui[1], Thales Bertaglia[1], Taylor Annabell[1],
Catalina Goanta[1], Tjomme Dooper[2], and Gerasimos Spanakis[3]

[1]Utrecht University, The Netherlands
[2]Stichting Reclame Code, The Netherlands
[3]Maastricht University, The Netherlands

Abstract

The rise of influencer marketing has blurred boundaries between organic content and sponsored content, making the enforcement of legal rules relating to transparency challenging. Effective regulation requires applying legal knowledge with a clear purpose and reason, yet current detection methods of undisclosed sponsored content generally lack legal grounding or operate as opaque "black boxes." Using 1,143 Instagram posts, we compare *gpt-5-nano* and *gemini-2.5-flash-lite* under three prompting strategies with controlled levels of legal knowledge provided. Both models perform strongly in classifying content as sponsored or not (F1 up to 0.93), though performance drops by over 10 points on ambiguous cases. We further develop a taxonomy of reasoning errors, showing frequent citation omissions (28.57%), unclear references (20.71%), and hidden ads exhibiting the highest miscue rate (28.57%). While adding regulatory text to the prompt improves explanation quality, it does not consistently improve detection accuracy. The contribution of this paper is threefold. First, it makes a novel addition to regulatory compliance technology by providing a taxonomy of common errors in LLM-generated legal reasoning to evaluate whether automated moderation is not only accurate but also legally robust, thereby advancing the transparent detection of influencer marketing content. Second, it features an original dataset of LLM explanations annotated by two students who were trained in influencer marketing law. Third, it combines quantitative and qualitative evaluation strategies for LLM explanations and critically reflects on how these findings can support advertising regulatory bodies in automating moderation processes on a solid legal foundation.

1 Introduction and background

The rapid rise of social media has made influencer marketing a central strategy for brands seeking to shape followers' purchasing decisions through influencers' reach and credibility (De Veirman et al., 2017). While effective at enhancing trust and engagement, this strategy is often opaque, as influencers generally avoid disclosures to maintain authenticity or protect follower engagement. Consequently, sponsored content is frequently hidden or inadequately disclosed (Ershov and Mitchell, 2020), limiting the consumers' ability to recognise advertising[1] and making regulatory oversight difficult.

Distinguishing ads from organic posts can be ambiguous (Figure 1); tagged brands may signal sponsorship or merely personal preference. Even with close scrutiny, regulators can misjudge cases, risking unfair penalties for legitimate influencers and causing complaints, as seen in (Code, 2023c) before the Dutch self-regulatory body *Stichting Reclame Code (SRC)*[2], where an independent jury justified its decision using legal reasoning.

The lack of transparency in influencer marketing is the largest issue consistently identified by self-regulatory bodies (Code, 2025; Practice, 2025; Almed, 2024). Self-regulators are industry organisations that make private rules for businesses. The main challenge for such bodies trying to measure compliance with their own rules in practice is the sheer amount of social media posts that can potentially contain commercial content. The fact that social media platforms do not allow anyone to thoroughly search their databases further complicates the enforcement of transparency standards. For practitioners, separating organic content from ads is the first step in assessing the compliance of influencer marketing with advertising law and self-regulatory codes. This is a laborious process that requires experts to spend their time viewing social media posts that might not contain any advertising.

[1]This paper uses the terms advertising, sponsored content (posts), and ads interchangeably

[2]https://www.reclamecode.nl/over-de-src/over-de-src/

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 157–171

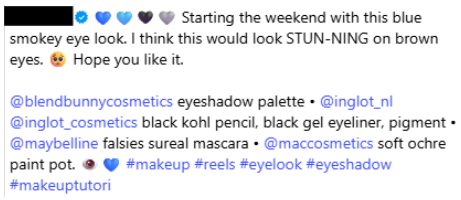

Figure 1: Example of an Instagram post that may be sponsored due to the presence of tagged brands.

Commercially available software platforms aid in this process by using keyword filters[3], which are usually not accurate enough to eliminate all organic posts from a sample.

In response to these challenges, computational research has sought to automate the detection of undisclosed ads (Zarei et al., 2020; Kim et al., 2021; Martins et al., 2022; Mathur et al., 2018; Bertaglia et al., 2023, 2024), but current methods face two limitations: (1) they often lack a solid legal foundation, exposing regulators to pushback in relation to their decisions, and (2) they prioritise accuracy over explanation (Rogers et al., 2023). Without a reasoning process, detection systems risk crossing boundaries that may conflict with free speech or other protected interests (Huang, 2025).

Large Language Models (LLMs) may offer promising ways to address both of these gaps. They can be prompted to reference relevant legal rules and provide explanations for their outputs, which makes the results more transparent and easier to interpret (Louis et al., 2024). At the same time, LLMs are prone to errors (e.g. hallucinations, weak grounding (Dahl et al., 2024; Bang et al., 2025)). This paper explores the potential of LLMs for detecting undisclosed influencer marketing by examining how they identify hidden advertising and evaluating the quality of their accompanying legal reasoning. Our main contributions are (1) a taxonomy of common errors in legal reasoning generated by LLM, extending previous research to a complex domain-specific context, namely, the detection of undisclosed advertisements on social networks, which also serves as a broader example of automatic compliance monitoring; (2) an original dataset of LLM explanations annotated by two students who were trained in influencer marketing law; (3) a demonstration of quantitative and qualitative evaluation strategies for LLM explanations and critically reflects on how these findings can support advertising regulatory bodies in automating moderation processes on a solid legal foundation. In general, our multidisciplinary approach, combining legal expertise with computer science, advances research on sponsored content detection and offers practical insights directly applicable to the industry. Finally, we make our material (data, code, annotation results) available.[4]

[3]For example, https://www.influencermonitor.com/

2 Related Work

2.1 NLP and sponsored content detection

Recent advances in detecting hidden advertisements on social media leverage both rule-based and machine learning approaches (Gui et al., 2025; Bertaglia et al., 2025). Rule-based methods detect explicit cues such as coupon codes or campaign hashtags like '#ad' with high precision (Santos Rodrigues et al., 2021; Swart et al., 2020), but struggle with implicit or unconventional disclosures. Machine learning methods, in contrast, capture complex, context-dependent patterns from annotated datasets. For example, Kim et al. (2021) combined textual, visual, and social network features to improve detection, Zarei et al. (2020) identified a notable share of undisclosed Instagram promotions, and Kok-Shun and Chan (2025) used GPT-4o to detect sponsored YouTube segments with high accuracy. Despite these gains, a shared limitation is that such models largely operate as black boxes, producing accurate predictions without interpretable reasoning.

2.2 LLM and legal texts

Parallel to advances in sponsored content detection, research has explored the ability of LLMs to process legal text across tasks such as legal question answering (Yuan et al., 2024), judgment prediction (Medvedeva and Mcbride, 2023; Chalkidis et al., 2022), contract review (Hendrycks et al., 2021), and legal reasoning (Guha et al., 2023), with reviews summarising tasks, datasets, methods, and challenges (Katz et al., 2023; Ariai and Demartini, 2025). General-purpose LLMs like GPT-4 and Claude perform well only after fine-tuning on legal examples (Blair-Stanek et al., 2024), motivating benchmarks that consolidate legal tasks into unified evaluation frameworks (Guha et al., 2023; Fei et al., 2024) or building more interpretable legal question answering models using a Retrieval-Augmented Generation (RAG) approach (Louis et al., 2024).

[4]https://github.com/HaoyangGui/Evaluating-LLM-Generated-Legal-Explanations

Related work has also examined LLM reasoning in legal adjacent domains, where the study is not only on the legal text but its application on real-world, real-user data, such as policy interpretation (Pałka et al., 2025; Palla et al., 2025) and content moderation (Kolla et al., 2024), highlighting both the potential and challenges of applying LLMs to specialised, rule-governed contexts similar to law.

2.3 Evaluating LLM output with legal knowledge

Recent research has moved beyond measuring the raw accuracy of LLMs in legal and policy-related tasks to evaluating the trustworthiness and quality of their explanations (Huang, 2025; Zhang et al., 2024; Calderon et al., 2025). Across domains such as legal reasoning (Kang et al., 2025; Mishra et al., 2025), policy enforcement (Pałka et al., 2025), and content moderation (Kolla et al., 2024), a key challenge is how to systematically assess LLM outputs in relation to legal knowledge. Despite this growing interest, progress is hindered by the lack of specific datasets that provide legally-informed annotations, which are critical for accurate benchmarking and systematically assessing both classification performance and the quality of generated legal reasoning.

Traditional evaluation metrics, such as accuracy, F1 score, and correlation, provide a baseline to assess classification performance (Bavaresco et al., 2025; Ashktorab et al., 2025; Tan et al., 2025; Trautmann et al., 2024), but they fail to capture LLMs' ability to understand context and nuance (Huang, 2025). Some studies incorporate lexical and semantic similarity (Vats et al., 2023), while broader computational metrics examine conflict rates among LLM annotators (Wang et al., 2024), plausibility and faithfulness of explanations (Shailya et al., 2025), groundedness (Trautmann et al., 2024), and stability (Blair-Stanek and Durme, 2025), often combined with statistical agreement with human experts (Chiang and Lee, 2023; Calderon et al., 2025).

Recognising that neither automated nor human judgments are perfectly accurate, recent work emphasises transparency in LLM-generated output, assessing qualities such as consistency, coherence, and informational richness (Golovneva et al., 2023; Prasad et al., 2023; Patel et al., 2024), alongside manually identifying reasoning errors (Li et al., 2023; Tyen et al., 2024; Mishra et al., 2025). For instance, Mishra et al. (2025) develops an error taxonomy for legal reasoning and methods to automate error detection. Collectively, these studies highlight that while LLMs show promise for legal and self-regulatory tasks, their out-of-the-box performance is limited, and fine-tuning is often required. Crucially, prior research has not extended these evaluation frameworks to complex, domain-specific contexts, such as legal interpretation in detecting undisclosed advertisements on social media, which is a key gap in compliance detection.

3 Study design and methodology

In this study, we evaluate how different LLMs classify influencer content and produce legal reasoning to justify their identification of advertising in the Dutch context. To this end, we created a dataset consisting of three types of content: disclosed advertisements, hidden advertisements, and organic posts (details are provided in the following section). The dataset is first fed into three different LLMs under three prompting strategies. Each model produces two outputs: (1) a binary classification indicating whether the post constitutes an advertisement, and (2) an accompanying explanation with legal reasoning to justify the decision. Then, for all posts and each type of content, we use two methods to examine the outputs:

Quantitative evaluation: We assess advertisement/organic content classification performance using standard classification metrics. This enables performance comparisons both within and across categories, and allows us to select the two best-performing models to proceed to the next step. Limiting further evaluation to these top-performing models helps avoid redundant comparisons and streamlines the analysis process. As a baseline, we use a TF-IDF (unigrams, bigrams) representation combined with logistic regression, employing an 80:20 train-test split. We did not include other deep learning models, such as BERT, as prior work suggests that they perform even worse in this context (Bertaglia et al., 2023).

Qualitative evaluation: We manually select balanced representative cases from each content type. Research assistants review the explanations by rating their helpfulness and annotating error types. This reveals systematic patterns linking specific errors to content types and prompting strategies. We also provide a case analysis, where a senior legal researcher reflects on the textual quality of a selection of outputs.

3.1 Dataset

The dataset used in this study originates from Gui et al. (2024) and comprises 300,199 posts by influencers registered in the Dutch Video-Uploader Registry [5]. For the purposes of this research, we focus exclusively on Instagram as the platform of interest and restrict our analysis to posts written in English. In line with the standards established by Gui et al. (2024), we adopt the same criteria for identifying sponsorship disclosures. Specifically, we only include posts with so-called 'green disclosures' (legally sufficient disclosed advertisement), which meet the legal requirements set out in the Dutch Advertising Code, resulting in 592 posts.

To construct a dataset for classification purposes, we then randomly sample an equal number of posts without green disclosures drawn from the same set of influencers (15 or the maximum number of posts by each), resulting in 551 posts. These posts may contain either sponsored content or not; therefore, three domain experts annotated these posts, distinguishing between hidden advertisements and organic content. The final labels are assigned through a two-step process: two domain experts (ann1 and ann2) must reach consensus, with any disagreements or uncertain cases referred to the third domain expert (ann3). Excluding 10.34% uncertain cases, annotators 1 and 2 achieve a 92.64% absolute agreement rate and 0.74 Krippendorff's Alpha, indicating substantive agreement.

The final dataset includes 1,143 English-language posts: 592 disclosed ads, 127 undisclosed ads, and 424 organic posts. To evaluate the ability of LLMs to detect hidden advertising, all explicit disclosure cues (such as #ad, etc.) are removed from the disclosed ads before model input. Table 1 provides a detailed description of the dataset, showing that organic posts tend to be shorter and include fewer hashtags and mentions. In contrast, sponsored posts are generally more similar to each other than to organic content, which increases the challenge of accurately distinguishing between these categories.

3.2 Models and prompts

We employ three prompting strategies, each with identical task instructions but varying in the degree of provided legal knowledge. By gradually reducing the amount of legal context, we aim to examine the extent to which LLMs rely on and apply legal knowledge when identifying advertisements. In all cases, each prompt instructs the LLM to determine whether a post is advertising and to provide a legal reasoning explanation. The three levels of legal knowledge are defined as follows:

- **Original codes with explanations**: This prompt incorporates the full regulatory text issued by *Stichting Reclame Code (SRC)*, a Dutch self-regulatory organisation that promotes responsible advertising in addition to legislation. This prompt includes the original regulation text and the corresponding explanations from the *General Section* and the special *Advertising Code Social Media & Influencer Marketing (RSM)*. This context is the most comprehensive form of legal knowledge based on text.

- **Original codes without explanations**: This prompt contains the same full regulatory text from the SRC as above, but omits the explanatory notes.

- **Names of the advertising codes only**: This prompt merely references the titles of the two codes (General Section and RSM), without including the substantive legal texts.

To ensure comparability, we designed a single base instruction prompt (shown in Appendix A), which was adapted for each strategy. This base prompt was validated and refined through manual inspection of sample cases and iterative discussions among the co-authors. Although this process resulted in minor differences in wording across the three strategies, the overall task structure and requirements remained consistent.

We evaluated the three prompting strategies using three different LLMs: *gemini-2.5-flash-lite, gpt-4.1-nano, gpt-5-nano*. We ran all experiments with a temperature setting of 1 and used default values for all remaining hyperparameters.

3.3 Explanation evaluation: error annotations

One of the objectives of this study is to examine the extent to which LLMs can comprehend legal knowledge and apply it to justify their decisions through legal reasoning. To assess the quality of the explanations produced by the models, we define seven common error categories: (e1) *Wrong interpretation of legal citations*, (e2) *No citation*, (e3) *Citation is not clear*, (e4) *Hallucinations on the legal*

[5] https://www.cvdm.nl/registers/

	Disclosed	Organic	Undisclosed
Posts	592	424	127
Tokens (mean ± std)	51.67 ± 47.93	26.45 ± 45.65	41.04 ± 56.97
Hashtags (mean ± std)	2.23 ± 3.53	2.91 ± 7.18	1.84 ± 4.11
Mentions (mean ± std)	1.36 ± 1.06	0.51 ± 2.38	1.90 ± 2.19
Posts with hashtag (%)	58.78	33.25	41.73
Posts with mention (%)	90.88	12.74	93.70

Table 1: Descriptive statistics for posts by category. Means and standard deviations (std) are reported for tokens, hashtags, and mentions. Posts with hashtags/mentions (%) show the percentage of posts that have hashtags or mentions.

Model	Prompting strategy	Precision	Recall	F1
logistic regression (TF-IDF)		0.85	0.91	0.88
gemini-2.5-flash-lite	no_article	0.91	**0.93**	0.92
gemini-2.5-flash-lite	article	0.92	**0.93**	**0.93**
gemini-2.5-flash-lite	article_explanation	0.92	0.92	0.92
gpt-4.1-nano	no_article	0.88	0.87	0.87
gpt-4.1-nano	article	0.87	0.83	0.85
gpt-4.1-nano	article_explanation	0.86	0.83	0.85
gpt-5-nano	no_article	0.94	0.91	0.92
gpt-5-nano	article	0.94	0.87	0.91
gpt-5-nano	article_explanation	**0.95**	0.86	0.90

Table 2: Comparison of performance across models and prompting strategies for the whole dataset in the task of advertisement identification.

citations, (e5) *Hallucinations on the content*, (e6) *Mistaken potential cues*, and (e7) *Reasoning results in opposite output*. Detailed descriptions and examples are provided in Table 4 (Appendix B).

Two research assistants with legal knowledge (annA and annB) rated the helpfulness of a subset of explanations and annotated the presence of these errors. Since LLM outputs vary widely in length and content, we only note whether an error is present in an explanation (note that one explanation might contain multiple errors). Before annotation, the assistants received training from domain experts and completed revisions after resolving any ambiguities.

The evaluation sample includes 60 randomly selected posts, evenly distributed across three types of content: 20 disclosed ads, 20 hidden ads, and 20 organic posts. For hidden ads and organic posts, we further divide the 20 examples into two groups based on the earlier sponsorship annotation stage: 10 posts with consensus labels from annotators ann1 and ann2, and 10 labelled solely by ann3 (no consensus reached by ann1 and ann2).

For the evaluation of the explanations, annotators A and B labelled 10 overlapping posts (in addition to 25 distinct posts each), achieving 89.29% absolute agreement and 0.37 Krippendorf's Alpha. As we compare different LLMs (*gpt-5-nano* and *gemini-2.5-flash-lite*) under three prompting strategies, each annotator evaluates 210 explanation units (35 posts × 2 models × 3 prompting strategies). Using these annotations, we analyse and discuss how explanation quality varies across models, prompting strategies, and different types of content.

3.4 Explanation evaluation: case analysis

We complement the evaluation of the explanations with a qualitative, expert-driven evaluation of the results. For this, one of the authors of this paper, a senior legal scholar with expertise in Dutch advertising law, was assigned a random set of four explanations pertaining to two posts from the article_explanation prompt, one of which involves a disclosed advertisement and the other an undisclosed advertisement. While these examples cannot capture every factor present in the dataset, this case

analysis provides insight into the recurring patterns that characterise each experimental setting.

4 Results

We first evaluate the classification performance of three LLMs under three prompting strategies across the entire dataset in a zero-shot setup (i.e., without fine-tuning). Based on these results, we select the two best-performing LLMs for subsequent tasks, which include evaluating classification performance on each type of content and examining the quality of their explanations.

4.1 Classification results

Table 2 presents classification performance on the full dataset of 1,143 posts across all experimental settings. Overall, the results indicate that all models achieve reasonable performance, but *gpt-4.1-nano* consistently underperforms on every metric, even worse than the baseline, with F1 scores ranging from 0.85 to 0.87. To streamline further analyses, we focus on *gpt-5-nano* (GPT) and *gemini-2.5-flash-lite* (Gemini).

Examining model-level performance, GPT achieves the highest precision (0.95 with the article_explanation prompt), while Gemini demonstrates stronger recall (0.93) and generally higher F1 scores (0.93). Interestingly, the prompting strategy that incorporates the most legal knowledge (article_explanation) does not always yield the best overall classification performance. For GPT, although article_explanation maximises precision, it reduces recall, resulting in the lowest F1 (0.90). Similarly, for Gemini, the highest recall (0.93) is achieved without explanations (article prompt), highlighting that more legal knowledge does not automatically translate into better classification outcomes. Differences across prompting strategies are relatively small, but this pattern suggests that LLMs' ability to apply legal knowledge may rely more on patterns learned during pretraining rather than the provided legal text.

Next, we focused on 95 ambiguous posts where annotators (ann1 and ann2) disagreed or expressed uncertainty in the advertisement annotation procedure (section 3.1). As expected, overall performance dropped significantly, with F1 scores falling by over 10 percentage points compared to the full dataset. The baseline model exhibited an even steeper decline, exceeding a 30-point reduction. GPT shows high precision (0.80 with no_article prompt) but suffers from lower recall, whereas Gemini maintains stronger recall and balanced F1 scores (0.80), consistent with its relative strengths in the full dataset. Notably, no prompting strategy equipped with explanations consistently outperforms others, reinforcing the observation that adding explicit legal text does not guarantee improved performance, particularly on ambiguous or borderline cases. Detailed results are provided in Table 5 (Appendix B).

Zooming in on the results by types of content, Gemini performs better on disclosed and hidden ads (0.94 and 0.93), whereas GPT performs better on organic content (0.92). GPT's performance on hidden ads remains notably weaker, even weaker than the baseline model, suggesting that its precision-oriented strengths do not extend to detecting subtle or undisclosed advertising cues. Prompting strategies show no consistent pattern: for Gemini, 'article' prompts perform best overall, while 'no_article' prompts slightly lead on disclosed and hidden ads; for GPT, 'no_article' prompts dominate on disclosed and hidden ads, whereas legal-knowledge prompts are better for organic content. A more granular breakdown of accuracy by content type, model, and prompting strategy can be found in Table 6 (Appendix B).

4.2 Evaluation of explanations

To assess the quality of LLM-generated legal explanations, we consider two complementary dimensions: (1) their perceived helpfulness to annotators, and (2) the types and frequencies of errors they contain.

Helpfulness and errors by models and prompting strategies We begin by analysing the errors in the explanations as described above. The last row in Table 3 shows the percentage of error types observed in LLM-generated explanations across all annotated posts. The most frequent error is e2 (No citation, 28.57%), followed by e3 (Unclear citation, 20.71%), indicating that LLMs often attempt but fail to provide explicit legal references. Less common errors include e1 (Wrong interpretation, 8.57%), e6 (Mistaken cues, 7.38%), e4 (Hallucinated citations, 2.62%), and e5 (Hallucinated content, 2.38%), while e7 (Contradictory reasoning, 0.24%) is rare. These patterns raise a key question: do models genuinely understand legal content or simply produce superficially plausible explanations?

	Model	Variant	Helpfulness score	e1 (%)	e2 (%)	e3 (%)	e4 (%)	e5 (%)	e6 (%)	e7 (%)
0	gemini-2.5-flash-lite	no_article	3.31 ± 0.84	**17.14**	**81.43**	25.71	5.71	2.86	10.00	0.00
1	gemini-2.5-flash-lite	article	$\mathbf{4.37 \pm 0.89}$	5.71	5.71	7.14	0.00	**7.14**	**12.86**	1.43
2	gemini-2.5-flash-lite	article_explanation	$\mathbf{4.37 \pm 0.75}$	7.14	2.86	4.29	0.00	1.43	10.00	0.00
3	gpt-5-nano	no_article	3.29 ± 0.93	2.86	78.57	**35.71**	**7.14**	1.43	2.86	0.00
4	gpt-5-nano	article	4.20 ± 1.10	11.43	1.43	20.00	1.43	1.43	4.29	0.00
5	gpt-5-nano	article_explanation	4.13 ± 0.99	7.14	1.43	31.43	1.43	0.00	4.29	0.00
Total				8.57	28.57	20.71	2.62	2.38	7.38	0.24

Table 3: Helpfulness score and error rates across models and prompting strategies for all types of content. Error types: (e1) *Wrong interpretation of legal citations*, (e2) *No citation*, (e3) *Citation is not clear*, (e4) *Hallucinations on the legal citations*, (e5) *Hallucinations on the content*, (e6) *Mistaken potential cues*, and (e7) *Reasoning results in opposite output*. For each error, the value shown is the proportion of posts containing the corresponding error. The last row shows the percentage of each error across the whole dataset.

Table 3 also presents a detailed assessment of LLM explanations in terms of both their perceived helpfulness and the percentage of data that contains corresponding types of errors across different models and prompting strategies. Helpfulness scores (1–5 scale) show Gemini with article_explanation performs best (4.37 ± 0.75), followed by GPT with article prompts (4.20 ± 1.10). No_article variants for both models achieve the lowest scores, indicating that legal input, especially when combined with explanations, improves perceived reasoning quality.

Critical citation errors (e2, e3) dominate no_article prompts: 81.43% and 25.71% for Gemini, 78.57% and 35.71% for GPT. Even with legal input, GPT still shows notable e1/e3 rates (7.14%/31.43% for article_explanation, 11.43%/20% for article), whereas Gemini's rates are lower (7.14%/4.29% for article_explanation, 5.71%/7.14% for article). However, Gemini exhibits higher e6 under article prompts (12.86%), showing that legal text alone does not guarantee accurate interpretation. In contrast, hallucinations (e4, e5) remain rare but concerning. Nearly all e4 cases occur in no_article prompts (5.71% Gemini, 7.14% GPT), where models fabricate citations due to missing legal context.

Errors by content type We also analyse errors by content type. Disclosed ads show the lowest rates for most errors, except for some e1–e3 cases in no_article variants. With legal context, hallucinations (e4, e5) are virtually absent, indicating that models rarely fabricate legal citations or misrepresent content when sufficient context is given. Detailed results can be found in Table 7 (Appendix B).

Undisclosed ads exhibit the highest e6 rate (28.57%) and notable e3 errors, with e4 and e5 appearing more often than in other categories. These patterns reflect the difficulty of detecting subtle promotions, where models must infer intent from indirect cues and often misidentify which signals indicate sponsorship.

Organic content shows comparatively higher e2 (No citation) and e3 (Unclear citation) errors, especially under no_article prompts, suggesting that models sometimes false legal reasoning without a real basis. Moderate e6 levels further indicate a tendency to overfit and misread ordinary content as promotional, highlighting the inherent ambiguity of influencer posts.

Case analysis: examining legal reasoning From a legal perspective, the task is simple, albeit domain-specific. Legal explanations follow an innate structure, due to the relevance of logic for legal argumentation (Bench-Capon et al., 2009; Lind, 2014). The task at hand involves identifying whether a post constitutes advertising. Our case analysis reveals that neither model was able to generate a cohesive, well-structured legal explanation. The model outputs an amalgam of statements, which is comparable to a rather poorly performing first-year law student. To be considered a basic but complete legal analysis, the output needed better performance in terms of selecting relevant provisions and in terms of structure.

In terms of provisions, according to the Dutch Advertising Code, which is industry self-regulation in the Netherlands, the starting point in determining whether something is advertising is that it has to fulfil all the conditions of Article 1 in Code (2023a) and Article 2. (c, d, e) in Code (2023b). While some dimensions of this definition cannot be analysed without additional facts (e.g., the relationship between an advertiser and a third party), some very concrete conditions should have been considered in an explanation, such as whether a post on Insta-

gram is public, whether the promotion of goods or services is direct or indirect, or whether the post consists in an idea, a good or a service. The four explanations in our case analysis mention Article 1, but there is generally a lack of systematic tackling of the conditions. In addition, the models seem to try to select and discuss many other articles, sometimes irrelevant (e.g., GPT mentioning Article 8.4). In terms of structure, there is no acknowledgement that a legal analysis is a demonstration that needs to be built according to some form of structure.

Generally, such a structure will differ from country to country or across fields of legal theory and practice; an inherent and easily detectable logic is necessary. All four explanations seem to provide some sort of conclusion, whether explicitly recognised as such or not, but the conclusion sometimes makes logical jumps, or it is a demonstration of conditions which are not relevant. Based on these factors, the explanations might seem, at first sight, to have relevance and accuracy, but upon closer examination, they are either chaotic, incomplete, or simply inaccurate.

5 Discussion and conclusion

This study examined how large language models (LLMs) can be applied to detect undisclosed advertising on social media while providing legal reasoning. Unlike prior research, which focused almost exclusively on classification accuracy, our work systematically evaluates both the quality of classification and the legal soundness of LLM explanations. This dual lens highlights critical gaps in current practice and suggests pathways toward more transparent and accountable automated moderation systems.

Starting from the classification task, both *gpt-5-nano* and *gemini-2.5-flash-lite* achieve high overall accuracy in identifying advertising content, but model choice strongly influences both classification strength and error profile: Gemini is more effective for recall-oriented tasks such as detecting hidden ads, whereas GPT excels in precision. Notably, LLMs are not always superior to simple baselines in overall classification performance; however, they perform better in challenging cases. Similar patterns of strength appear in the 95 ambiguous posts, where annotators (ann1 and ann2) disagreed or expressed uncertainty in the advertisement annotation procedure (section 3.1). Examining the content further, these patterns of ambiguity align with previous findings, which attribute annotator disagreement to both data-related factors (e.g., various language features, uncertainty in sentence meaning), and annotator-related factors (e.g. various language features, uncertainty in sentence meaning) Jiang and de Marneffe (2022); Plank (2022-12); Xu et al. (2023-12). These intrinsic complexities pose challenges for LLMs, contributing to lower performance in ambiguous contexts.

Moreover, increasing the amount of embedded legal text does not consistently improve the classification outcomes. While prompts containing full regulatory codes and explanations raise the perceived helpfulness of LLM reasoning (e.g., Gemini article_explanation reaching 4.37 ± 0.75 versus 4.20 ± 1.10 for GPT), they do not guarantee better moderation outcomes. This indicates that current LLMs do not simply "read and apply" legal norms; instead, they rely heavily on internal heuristics and contextual associations. In practice, this means that LLMs are already capable of recognising different forms of advertising because promotional language and stylistic cues are strongly represented in their training data. Cues indicating sponsorship, patterns of product placement, or persuasive rhetorical devices can often be detected without direct reference to regulatory codes. In this sense, the models' performance may reflect an underlying competence in identifying pragmatic markers of advertising, rather than understanding and applying legal knowledge as a content moderator.

The explanation analysis further reveals systematic weaknesses. Citation-related errors, missing (e2, 28.57%), unclear (e3, 20.71%), or wrong interpretations (e1, 8.57%), dominate across settings, particularly when no legal text is provided. Even when legal sources are available, models often select irrelevant provisions or fail to structure reasoning in a way consistent with basic legal methodology. More severe hallucinations of legal citations (e4, 2.62%) and content (e5, 2.38%) are rare but concentrated in no_article prompts, where GPT and Gemini fabricated legal references at 7.14% and 5.71%, respectively. These patterns suggest that LLMs tend to approximate legal reasoning rather than reliably apply normative rules, which essentially means that they fail to 'read, understand, and apply.'

A closer look by content type further illuminates these limitations. Undisclosed ads produce the highest rate of misidentified cues (e6, 28.57%), showing that LLMs frequently mistake ordinary or

ambiguous content for sponsored posts. In contrast, disclosed ads show almost no hallucinations when legal text is provided, indicating that straightforward content allows LLMs to stabilise their reasoning more reliably. Together with the case analysis carried out, these findings confirm that although LLMs can approximate legal reasoning, they are far from delivering rigorous justifications akin to an expert with domain-specific knowledge.

These findings have two broad implications for moderation. First, they demonstrate that high classification accuracy does not ensure trustworthy enforcement. An LLM that labels a post correctly but cites irrelevant or fabricated legal provisions cannot satisfy procedural fairness standards. Second, explanation quality varies systematically by content type and prompting strategy, meaning that moderation pipelines cannot rely on a one-size-fits-all approach. Platforms using LLMs for detection must pair performance metrics with legal-reasoning audits to ensure that decisions are not only correct but also defensible. In practice, this means building tools that flag cases with high-risk errors (e.g., e4/e5 hallucinations) for human review and calibrating models to reduce over-classification in ambiguous contexts.

These findings also connect to broader debates on moderation with LLMs. As Goanta et al. (2023) argues, NLP research must be situated within regulatory studies to avoid regulatory capture and to bridge the "pacing gap" between technological innovation and legal adaptation. Our results reflect this concern: models that appear accurate can still misapply or fabricate legal norms, undermining the legitimacy of enforcement. Treating moderation as a purely technical task risks obscuring the regulatory standards it is supposed to serve; instead, explanation quality and legal soundness must be foregrounded alongside accuracy. At the same time, our taxonomy of explanation errors resonates with emerging moderation research that highlights the concerns of LLMs as moderators. Yin et al. (2025) demonstrates that binary safe/unsafe labels miss important gradations of harm. Similarly, in our research, not all explanation errors are equally harmful: vague reasoning may be tolerable, but fabricated citations or misapplied provisions threaten procedural fairness. Integrating severity-sensitive auditing into compliance monitoring would thus allow regulators to triage high-risk cases while ensuring that enforcement remains both effective and legitimate.

The main contribution of this paper is to integrate the quality of legal reasoning in the evaluation of influencer marketing detection systems. By developing a taxonomy of LLM explanation errors and showing how these patterns vary by model, prompting strategy, and content type, we provide an actionable framework for regulators and platform designers. Instead of treating LLM outputs as opaque predictions, our study demonstrates how to assess whether automated moderation is not only accurate but also legitimate. This is particularly valuable for self-regulatory bodies such as *Stichting Reclame Code (SRC)*, which must justify enforcement decisions in legal terms rather than through statistical metrics alone. More broadly, our multidisciplinary approach, combining computational evaluation with legal analysis, offers a blueprint for building moderation systems that are transparent, explainable, and aligned with rule-of-law principles rather than black-box heuristics.

Limitations

Our dataset focuses solely on textual content, excluding visual or multimodal signals that frequently convey sponsorship. Human annotation also entails subjectivity, especially for borderline cases where even experts disagree. Moreover, the study relies on off-the-shelf LLMs without fine-tuning, meaning performance could improve with domain-specific adaptation.

Acknowledgments

This research has been supported by funding from the ERC Starting Grant HUMANads (ERC-2021-StG No 101041824). We also thank Isolde Torres and Giulio Bernasconi for their valuable assistance with this research.

References

IAP Almed. 2024. Monitoring Transparency and Influencer Marketing: Beauty, fashion, family and finance.

Farid Ariai and Gianluca Demartini. 2025. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *Preprint*, arXiv:2410.21306.

Zahra Ashktorab, Michael Desmond, Qian Pan, James M. Johnson, Martin Santillan Cooper, Elizabeth M. Daly, Rahul Nair, Tejaswini Pedapati, Hyo Jin Do, and Werner Geyer. 2025. Aligning human and LLM judgments: Insights from

EvalAssist on task-specific evaluations and AI-assisted assessment strategy preferences. *Preprint*, arXiv:2410.00873.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. *Preprint*, arXiv:2504.17550.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255. Association for Computational Linguistics.

Trevor Bench-Capon, Henry Prakken, and Giovanni Sartor. 2009. Argumentation in legal reasoning. In Guillermo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 363–382. Springer US.

Thales Bertaglia, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2025. Influencer self-disclosure practices on Instagram: A multi-country longitudinal study. *Online Social Networks and Media*, 45:100298.

Thales Bertaglia, Lily Heisig, Rishabh Kaushal, and Adriana Iamnitchi. 2024. Instasynth: Opportunities and challenges in generating synthetic instagram data with chatgpt for sponsored content detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 139–151.

Thales Bertaglia, Stefan Huber, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2023. Closing the loop: Testing chatgpt to generate model explanations to improve human labelling of sponsored content on social media. In *World Conference on Explainable Artificial Intelligence*, pages 198–213. Springer.

Andrew Blair-Stanek and Benjamin Van Durme. 2025. LLMs provide unstable answers to legal questions. *Preprint*, arXiv:2502.05196.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2024. BLT: Can large language models handle basic legal text? In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 216–232. Association for Computational Linguistics.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. *Preprint*, arXiv:2501.10970.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631. Association for Computational Linguistics.

Stichting Reclame Code. 2023a. General - stichting reclame code.

Stichting Reclame Code. 2023b. Special advertising codes - advertising code foundation.

Stichting Reclame Code. 2023c. Statement - advertising code foundation. Stichting Reclame Code.

Stichting Reclame Code. 2025. Certification works: Influencer violations halved after e-learning - Advertising Code Foundation.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Marijke De Veirman, Veroline Cauberghe, and Liselot Hudders. 2017. Marketing through Instagram influencers: The impact of number of followers and product divergence on brand attitude. *International Journal of Advertising*, 36(5):798–828.

Daniel Ershov and Matthew Mitchell. 2020. The effects of influencer advertising disclosure regulations: Evidence from instagram. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, pages 73–74. Association for Computing Machinery.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962. Association for Computational Linguistics.

Catalina Goanta, Nikolaos Aletras, Ilias Chalkidis, Sofia Ranchordás, and Gerasimos Spanakis. 2023. Regulation and NLP (RegNLP): Taming large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8712–8724. Association for Computational Linguistics.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. *Preprint*, arXiv:2212.07919.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Haoyang Gui, Thales Bertaglia, Catalina Goanta, Sybe de Vries, and Gerasimos Spanakis. 2024. Across platforms and languages: Dutch influencers and legal disclosures on instagram, YouTube and TikTok. In *Social Networks Analysis and Mining: 16th International Conference, ASONAM 2024, Rende, Italy, September 2–5, 2024, Proceedings, Part III*, pages 3–12. Springer-Verlag.

Haoyang Gui, Thales Bertaglia, Catalina Goanta, and Gerasimos Spanakis. 2025. Computational studies in influencer marketing: A systematic literature review. *Preprint*, arXiv:2506.14602.

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

Tao Huang. 2025. Content moderation by LLM: From accuracy to legitimacy. *Preprint*, arXiv:2409.03219.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Zhuang Li, and Adnan Trakic. 2025. Automating IRAC analysis in malaysian contract law using a semi-structured knowledge base. *Artificial Intelligence and Law*.

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

Seungbae Kim, Jyun-Yu Jiang, and Wei Wang. 2021. Discovering undisclosed paid partnership on social media via aspect-attentive sponsored post learning. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pages 319–327. Association for Computing Machinery.

Brice Valentin Kok-Shun and Johnny Chan. 2025. Leveraging ChatGPT for sponsored ad detection and keyword extraction in YouTube videos. *Preprint*, arXiv:2502.15102.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, pages 1–8. Association for Computing Machinery.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making large language models better reasoners with step-aware verifier. *Preprint*, arXiv:2206.02336.

Douglas Lind. 2014. The significance of logic for law. The National Judicial College.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models: 38th AAAI conference on artificial intelligence 2024. *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 38:22266–22275.

Emanuelle Azevedo Martins, Isadora Salles, Fabricio Benevenuto, and Olga Goussevskaia. 2022. Characterizing sponsored content in facebook and instagram. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, pages 52–63. Association for Computing Machinery.

Arunesh Mathur, Arvind Narayanan, and Marshini Chetty. 2018. Endorsements on social media: An empirical study of affiliate marketing disclosures on YouTube and pinterest. *Proc. ACM Hum.-Comput. Interact.*, 2.

Masha Medvedeva and Pauline Mcbride. 2023. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84. Association for Computational Linguistics.

Venkatesh Mishra, Bimsara Pathiraja, Mihir Parmar, Sat Chidananda, Jayanth Srinivasa, Gaowen Liu, Ali Payani, and Chitta Baral. 2025. Investigating the shortcomings of LLMs in step-by-step legal reasoning. *Preprint*, arXiv:2502.05675.

Konstantina Palla, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Andreas Damianou, Henrik Lindström, Dan Taber, and Mounia Lalmas. 2025. Policy-as-prompt: Rethinking content moderation in the age of large language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 840–854. Association for Computing Machinery.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. 2024. Multi-LogiEval: Towards evaluating multi-step logical reasoning ability of large language models. *Preprint*, arXiv:2406.17169.

Przemysław Pałka, Francesca Lagioia, Rūta Liepina, Marco Lippi, and Giovanni Sartor. 2025. Make privacy policies longer and appoint LLM readers. *Artificial Intelligence and Law*.

Barbara Plank. 2022-12. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682. Association for Computational Linguistics.

Advertising Standards Authority {\textbar} Committee of Advertising Practice. 2025. Influencer ad disclosure on social media: Instagram and TikTok report (2024).

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086. Association for Computational Linguistics.

Anna Rogers, Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Alexander Koller, Sasha Luccioni, Maarten Sap, Roy Schwartz, Noah A. Smith, and Emma Strubell. 2023. Closed AI models make bad baselines.

João P. Santos Rodrigues, Ana C. Munaro, and Emerson Cabrera Paraiso. 2021. Identifying sponsored content in YouTube using information extraction. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3075–3080.

Krithi Shailya, Shreya Rajpal, Gokul S. Krishnan, and Balaraman Ravindran. 2025. LExT: Towards evaluating trustworthiness of natural language explanations. *Preprint*, arXiv:2504.06227.

Michael Swart, Ylana Lopez, Arunesh Mathur, and Marshini Chetty. 2020. Is this an ad?: Automatically disclosing online endorsements on YouTube with AdIntuition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12. Association for Computing Machinery.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. JudgeBench: A benchmark for evaluating LLM-based judges. *Preprint*, arXiv:2410.12784.

Dietrich Trautmann, Natalia Ostapuk, Quentin Grail, Adrian Pol, Guglielmo Bonifazi, Shang Gao, and Martin Gajek. 2024. Measuring the groundedness of legal question-answering systems. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 176–186. Association for Computational Linguistics.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. *Preprint*, arXiv:2311.08516.

Shaurya Vats, Atharva Zope, Somsubhra De, Anurag Sharma, Upal Bhattacharya, Shubham Kumar Nigam, Shouvik Guha, Koustav Rudra, and Kripabandhu Ghosh. 2023. LLMs – the good, the bad or the indispensable?: A use case on legal statute prediction and legal judgment prediction on indian court cases. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12451–12474. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450. Association for Computational Linguistics.

Shanshan Xu, Santosh T.y.s.s, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2023-12. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9576. Association for Computational Linguistics.

Fan Yin, Philippe Laban, Xiangyu Peng, Yilun Zhou, Yixin Mao, Vaibhav Vats, Linnea Ross, Divyansh Agarwal, Caiming Xiong, and Chien-Sheng Wu. 2025. BingoGuard: LLM content moderation tools with risk levels. *Preprint*, arXiv:2503.06550.

Mingruo Yuan, Ben Kao, Tien-Hsuan Wu, Michael M. K. Cheung, Henry W. H. Chan, Anne S. Y. Cheung, Felix W. H. Chan, and Yongxi Chen. 2024. Bringing legal knowledge to the public by constructing a legal question bank using large-scale pre-trained language model. *Artificial Intelligence and Law*, 32(3):769–805.

Koosha Zarei, Damilola Ibosiola, Reza Farahbakhsh, Zafar Gilani, Kiran Garimella, Noël Crespi, and Gareth Tyson. 2020. Characterising and detecting sponsored influencer posts on instagram. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 327–331.

Ruizhe Zhang, Haitao Li, Yueyue Wu, Qingyao Ai, Yiqun Liu, Min Zhang, and Shaoping Ma. 2024. Evaluation ethics of LLMs in legal domain. arXiv.org.

A Prompt template

Identity

You are a legal expert, as well as a social media content moderator who is responsible for keeping monetised posts compliant with the advertisement disclosure rules.

Context

You are reviewing social media posts that are likely to be undisclosed ads. Your goal is to determine, under Dutch advertising law, whether the

post is in fact an advertisement – regardless of whether disclosure is present. Disclosed posts should still be classified as ads if they meet the criteria. The classification is based on the nature of the post, not solely the presence/absence of disclosure.

Task

You're given these social media posts. Based on your legal knowledge of Dutch advertising law, decide if this post is an advertisement. First, justify your decision step-by-step using legal and contextual reasoning, referring to the specific articles from the regulations, and making a legal argument.

Output format

Please provide the following outputs, in this order, strictly adhering to the instructions and avoiding verbosity:

<Justification> Output the detailed reasoning that directed your result. This must be the chain-of-thought style legal reasoning, grounded in the Dutch Advertising Code and the Advertising Code Social Media & Influencer Marketing.

<Is the post an advertisement> True (1)/False (0). Output strictly as 1 or 0.

Always decide the label only after completing the reasoning.

B Extra tables

Error Type	Description	Example
Wrong interpretation of legal citations	The LLM gives an argument based on certain articles, but that is not what the article means	There is no disclosure or clear indication that this is a promotional post or part of an advertising campaign (Articles 11 and 3 of Dutch Advertising Code). Article 3: Advertising may not be contrary to the general interest, public order, or morality.
No citation	Explanations don't include any legal citations	—
Citation is not clear	It cites multiple articles but didn't clearly map them out	There is no disclosure or clear indication that this is a promotional post or part of an advertising campaign (Articles 11 and 3 of the Dutch Advertising Code and RSM). It didn't name specifically which article is from which code.
Hallucinations on the legal citations	When the answer includes legal information that is not in the regulation	According to Article 7, but there is actually no Article 7. According to (Some random law that you can check on Google if it really exists).
Hallucinations on the content	Besides legal content, the answer includes content that doesn't exist, such as the brand name	The influencer cooperate with @Nike, but actually there is no mentioning of Nike at all in the original post.
Mistaken potential cues	Don't/Wrongly identify a clue as advertisements or advertisers	#fyp is not an ad cue, but LLM believe it is; @a friend, but recognises that as an advertiser. Find the potential clues (#Nike), but don't take them as the evidence.
Reasoning ends up opposite the output	The reasoning process is opposite to the final conclusion. It means trying to reason it as an ad, but the final conclusion said it is not	Is there a Relevant Relationship? – Yes, the post explicitly mentions collaboration with @thewoolmarkcompany, indicating a business relationship. This relationship influences the content, as the post promotes wool products, possibly as part of sponsored content. With explanations all like this, it still label the post as False (non-ad).

Table 4: Types of errors in LLM responses regarding advertising identification.

	Model	Prompting strategy	Precision	Recall	F1 Score
0	logistic regression (TF-IDF)		0.60	0.55	0.57
1	gemini-2.5-flash-lite	no_article	0.72	0.84	0.77
2	gemini-2.5-flash-lite	article	0.74	**0.88**	**0.80**
3	gemini-2.5-flash-lite	article_explanation	0.75	0.86	**0.80**
4	gpt-5-nano	no_article	**0.80**	0.79	**0.80**
5	gpt-5-nano	article	0.75	0.70	0.73
6	gpt-5-nano	article_explanation	0.76	0.61	0.68

Table 5: Comparison of performance across models and prompting strategies for the ambiguous cases in the task of advertisement identification

	Model	Prompting strategy	Category	Accuracy	Category	Accuracy	Category	Accuracy
0	logistic regression (TF-IDF)		Disclosed ads	0.92	Hidden ads	0.86	Organic	0.73
1	gemini-2.5-flash-lite	no_article	Disclosed ads	**0.94**	Hidden ads	0.91	Organic	0.86
2	gemini-2.5-flash-lite	article	Disclosed ads	**0.94**	Hidden ads	**0.93**	Organic	0.88
3	gemini-2.5-flash-lite	article_explanation	Disclosed ads	0.93	Hidden ads	0.91	Organic	0.88
4	gpt-5-nano	no_article	Disclosed ads	0.92	Hidden ads	0.87	Organic	0.90
5	gpt-5-nano	article	Disclosed ads	0.90	Hidden ads	0.79	Organic	**0.92**
6	gpt-5-nano	article_explanation	Disclosed ads	0.89	Hidden ads	0.76	Organic	**0.92**

Table 6: Accuracy by model, prompting strategy, and type of content.

	Model	Variant	Data Source	e1 (%)	e2 (%)	e3 (%)	e4 (%)	e5 (%)	e6 (%)	e7 (%)
0	gemini-2.5-flash-lite	article	disclosed_ads	0.00	4.35	4.35	0.00	0.00	0.00	0.00
1	gemini-2.5-flash-lite	article	organic	3.85	7.69	11.54	0.00	7.69	11.54	0.00
2	gemini-2.5-flash-lite	article	undisclosed_ads	14.29	4.76	4.76	0.00	14.29	28.57	4.76
3	gemini-2.5-flash-lite	article_explanation	disclosed_ads	0.00	4.35	0.00	0.00	0.00	0.00	0.00
4	gemini-2.5-flash-lite	article_explanation	organic	11.54	0.00	7.69	0.00	3.85	19.23	0.00
5	gemini-2.5-flash-lite	article_explanation	undisclosed_ads	9.52	4.76	4.76	0.00	0.00	9.52	0.00
6	gemini-2.5-flash-lite	no_article	disclosed_ads	17.39	78.26	26.09	8.70	0.00	0.00	0.00
7	gemini-2.5-flash-lite	no_article	organic	3.85	92.31	23.08	0.00	3.85	15.38	0.00
8	gemini-2.5-flash-lite	no_article	undisclosed_ads	33.33	71.43	28.57	9.52	4.76	14.29	0.00
9	gpt-5-nano	article	disclosed_ads	13.04	4.35	39.13	0.00	0.00	0.00	0.00
10	gpt-5-nano	article	organic	7.69	0.00	3.85	0.00	0.00	7.69	0.00
11	gpt-5-nano	article	undisclosed_ads	14.29	0.00	19.05	4.76	4.76	4.76	0.00
12	gpt-5-nano	article_explanation	disclosed_ads	4.35	0.00	39.13	0.00	0.00	0.00	0.00
13	gpt-5-nano	article_explanation	organic	7.69	0.00	26.92	3.85	0.00	7.69	0.00
14	gpt-5-nano	article_explanation	undisclosed_ads	9.52	4.76	28.57	0.00	0.00	4.76	0.00
15	gpt-5-nano	no_article	disclosed_ads	0.00	95.65	26.09	4.35	0.00	0.00	0.00
16	gpt-5-nano	no_article	organic	3.85	76.92	42.31	7.69	0.00	7.69	0.00
17	gpt-5-nano	no_article	undisclosed_ads	4.76	61.90	38.10	9.52	4.76	0.00	0.00

Table 7: Error percentages by model, prompting strategy, and data source. Each value represents the proportion of posts containing the corresponding error. (e1) *Wrong interpretation of legal citations*, (e2) *No citation*, (e3) *Citation is not clear*, (e4) *Hallucinations on the legal citations*, (e5) *Hallucinations on the content*, (e6) *Mistaken potential cues*, and (e7) *Reasoning results in opposite output*. Each value represents the proportion of posts exhibiting the corresponding error.