

ADCOMPLIANCE: Towards Automated Compliance Verification in Advertising Operations

Saki Sugiyama[†], Peinan Zhang[‡], Yusuke Sakai[†],
Sho Takase[‡], Hiroki Ouchi^{†‡}, Taro Watanabe[†]

[†]Nara Institute of Science and Technology (NAIST), [‡]CyberAgent AI Lab
sugiyama.saki.sx7@naist.ac.jp, {zhang_peinan, takase_sho}@cyberagent.co.jp
{sakai.yusuke.sr9, hiroki.ouchi, taro}@is.naist.jp

Abstract

Advertising operations (AdOps) include compliance review processes to ensure that advertisements adhere to applicable laws, platform policies, and industry guidelines. In particular, ads must avoid issues such as overclaiming effectiveness or making definitive assertions about uncertain outcomes. However, these reviews are typically conducted by experts, making them an operational bottleneck. In this paper, we propose the automated ad text compliance verification task and present our classifier-based system. We introduce ADCOMPLIANCE, a dataset consisting of real pre-submission ad texts annotated by experts. Our experimental results demonstrate practical classification performance, suggesting that non-compliant ads can be effectively filtered while maintaining a low false positive rate. These results support practical integration into AdOps workflows.

1 Introduction

Nowadays, online advertising has become one of the dominant sectors of the advertising industry. In particular, as illustrated in Figure 1, a major role of ad agencies is to create sponsored search advertisements, such as short ad texts displayed alongside search engine results, e.g., Google Search, under strict character-length constraints (Zhang et al., 2025). The online advertising market is expected to grow by billions of dollars over the next several years (Murakami et al., 2023). Furthermore, advertising media are expanding beyond traditional web search, with new channels embedding advertisements across a variety of digital environments (Xu et al., 2026), making ad delivery an increasingly important component of the ecosystem.

Ad text generation is a typical task in creative text generation and has been widely studied (Kamigaito et al., 2021, 2024; Murakami et al., 2025a,b; Chen et al., 2025; Zhou et al., 2025). However, as shown in Figure 2, practical advertising opera-

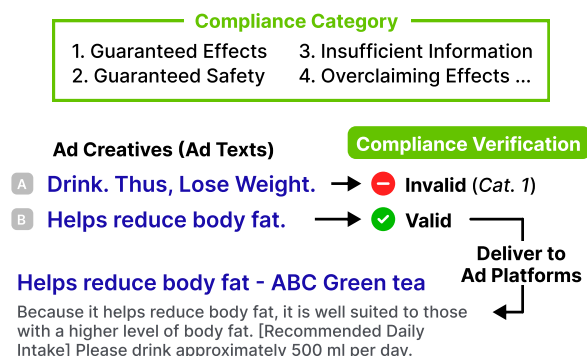


Figure 1: Illustration of compliance verification task. In the ad delivery process, ad text candidates must be verified to ensure compliance with applicable laws and advertising guidelines, such as avoiding overclaims or misleading statements. Currently, this verification is performed through manual review by experts. Our goal is to automate the compliance verification process, assisting experts while reducing human labor.

tions (AdOps) require not only generating advertisements but also evaluating them, for example, by assessing whether the content is appealing or selecting the most effective ad text under budget constraints. In practice, evaluating generated ad texts is often more important than generating them. Although recent studies (Zhang et al., 2025; Mita et al., 2024; Inoue et al., 2024; Kato et al., 2024; Murakami et al., 2022; Otake et al., 2025) have proposed application-oriented benchmarks that better reflect real advertising workflows of quality evaluation, these frameworks remain limited since they do not adequately cover the comprehensive review process required for practical ad submission.

One important remaining aspect is **compliance verification**. In an effort to make products appear more attractive, advertisements may contain exaggerated claims or overly definitive statements about their effectiveness. Such expressions can mislead consumers and violate regulations against deceptive advertising. As a result, many countries,

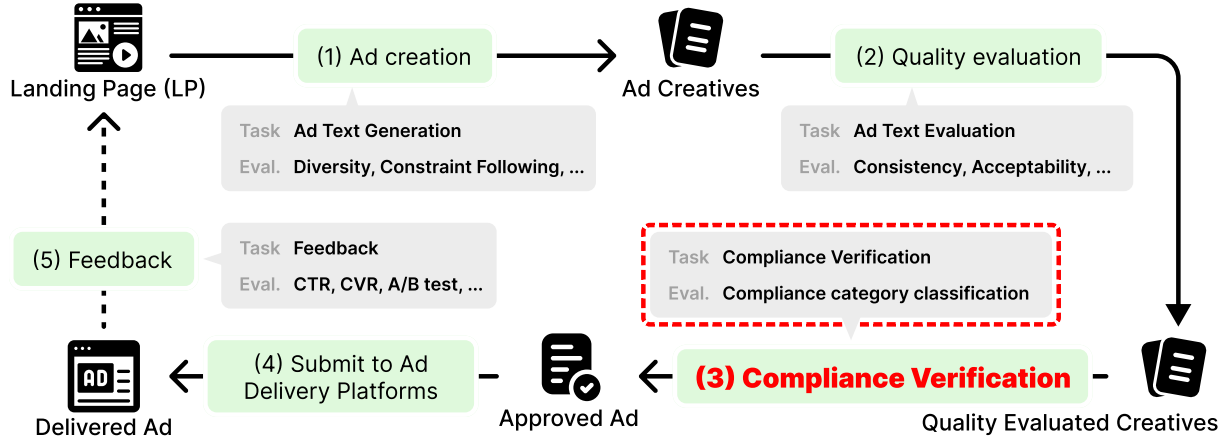


Figure 2: Overview of the advertising operations (AdOps) process and the position of the compliance verification task targeted in this study. First, an advertiser provides a landing page (LP), the web page to be promoted. Based on the LP, (1) the ad agency generates multiple candidate ad texts, referred to as ad creatives. These creatives are then reviewed by experts for both (2) quality and (3) compliance, including adherence to regional laws and advertising guidelines. After this review, (4) the approved ads are submitted to ad delivery platforms and served through channels such as sponsored search ads. (5) Performance feedback, such as click-through rates, is subsequently used to refine the LP or generate new ad creatives in an iterative cycle. While ad creation and quality evaluation have been formulated as well-defined tasks and are increasingly automated, compliance verification has received relatively little attention due to the specialized expertise required and still relies heavily on manual expert review.

advertising platforms, and industry alliances impose legal and policy restrictions on such claims¹. Submission of non-compliant advertisements can lead to consumer incidents and litigation risks. It can also damage an advertiser’s credibility, delay ad delivery, and, if repeated, even result in account suspension or bans, making compliance verification a critical step in the AdOps pipeline. However, this verification is still performed almost entirely through manual expert review, placing a substantial burden on human reviewers that continues to grow with increasing advertisement volume.

To address this challenge, we propose an automated ad text compliance verification task and introduce a compliance verification system for the AdOps workflow to assist expert reviewers and reduce manual workload, as illustrated in Figure 1. To support the development of such systems, we also introduce **AdCompliance**, a dataset of real in-house Japanese pre-submission advertisement texts annotated by experts. Experimental results show that our Japanese ModernBERT (Tsukagoshi et al., 2025; Sugiura et al., 2025; Warner et al., 2025)-based verification system can effectively filter out non-compliant advertisements while maintaining a very low false positive rate, rarely rejecting valid

ads. Furthermore, it successfully filters approximately 70% of invalid ads while accurately distinguishing most rejection reasons. By integrating this verification system into the AdOps workflow, human experts can focus primarily on borderline cases and final decision-making, thereby enabling a more efficient and scalable advertising process.

2 Background and Related Work

Ads Delivery Process. As shown in Figure 2, the real AdOps workflow delivers advertisements through a series of evaluation and verification stages applied to generated ad texts, i.e., **ad creatives**. Given a landing page (LP) provided by the advertiser, the ad agency first creates multiple candidate ad texts, referred to as ad creatives. These candidates undergo internal reviews, after which the advertisements expected to be the most effective are prioritized for delivery to users. Based on performance feedback, such as click-through rates, page views, traffic to the LP, and actual purchase outcomes, the ad texts are iteratively refined to improve advertising effectiveness (Zhang et al., 2025). In recent years, much research has focused on automating ad creative creation, i.e., ad text generation (Lei et al., 2022). As a result, ad text generation is becoming an increasingly mature task in natural language generation (Murakami et al., 2023). Moreover, Wang et al. (2025b,c) have ex-

¹E.g., Yahoo! Japan Ads Editorial Guidelines: https://ads-help.yahoo-net.jp/s/guideline-editorial?language=en_US (Accessed: June 10th, 2026).

panded the scope to multimodal advertising, including banner image generation. Nevertheless, these creatives still contain textual content, making textual information crucial even in multimodal settings. However, responsible AdOps ultimately requires expert approval before the advertisement deployment. As ad generation methods become increasingly practical and widely adopted, the burden of manual review continues to grow, making it a major bottleneck in the AdOps workflow.

Ads Evaluation. The evaluation of ad text generation has still primarily focused on benchmarking against human preferences, expert judgments, or actual advertising performance as gold labels, using metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), keyword matching, or task-specific evaluation metrics (Mita et al., 2024; Inoue et al., 2024; Kato et al., 2024). In addition, because diversity is an important objective in ad creative generation (Murakami et al., 2025a,b), self-evaluation metrics such as Self-BLEU (Zhu et al., 2018) are often employed to measure the diversity of generated advertisements. Beyond academic benchmarks, there is also a growing trend toward more practical evaluation settings that assess generated advertisements through click-through rates, conversion performance, and A/B testing in real advertising campaigns (Murakami et al., 2025b; Wang et al., 2025a). That said, before advertisements can be deployed in practice, they must satisfy a number of hidden criteria that are not captured by conventional benchmarks. As shown in Figure 2, ads are manually reviewed to ensure that they do not contain harmful or policy-violating content. These verification steps are indispensable components of real-world AdOps. AdTEC (Zhang et al., 2025) addresses this gap by introducing a primitive task to automate expert review in the AdOps workflow. However, it was designed primarily for ad quality evaluation.

Ads Compliance. In practice, AdOps also requires verification that advertisements comply with laws, government regulations², and industry alliance guidelines¹. As shown in Figure 1, advertisements containing definitive guarantees of effectiveness should not be delivered to customers because they may mislead them. Moreover, exaggerated safety or efficacy claims, such as sug-

gesting that taking more medicine will inevitably produce better results, can lead to litigation and even pose risks to human health. Such issues are particularly critical in domains such as pharmaceuticals³. Some studies, such as FAdR (Tang and Chen, 2014) and IAD (Liu et al., 2020), focus on detecting illegal advertisements that have already been published or collected from public sources, where the violations are relatively clear. In contrast, compliance verification in real-world AdOps workflows often requires judgments based on subtle nuances in ad texts. Therefore, integration into actual AdOps workflows requires verification of real pre-submission ad texts. In addition, experienced ad agencies often organize compliance knowledge into internal taxonomies and enforce standards that are even stricter than publicly available regulations in order to provide trustworthy advertising services. These standards reflect accumulated operational expertise and are designed to prevent future legal disputes and reputational risks. Consequently, compliance verification has become a major bottleneck in the AdOps workflow due to the difficulty of securing qualified experts, creating a strong demand for automated systems that assist human experts and reduce their workload.

3 AdCompliance

We propose an automated ad text verification task based on a basic compliance taxonomy to rapidly detect guideline violations in ad texts. We also introduce AD COMPLIANCE, a dataset constructed from annotation data collected through real-world advertising operations to support the development and evaluation of compliance verification systems.

3.1 Task Definition

The primary objective of the ad compliance verification task is to assign a binary label, **valid** or **invalid**, to ad creatives. As shown in Figure 1, valid ads are treated as positive instances, whereas invalid ads are treated as negative ones. The primary goal and evaluation criterion are to reduce the workload on human experts by automatically filtering out clearly invalid advertisements and forwarding only borderline cases for manual review, thereby improving the efficiency of the AdOps workflow.

²An example of Tokyo metropolitan government (In Japanese): <https://www.hokeniryo.metro.tokyo.lg.jp/anzen/iyaku/sonota/koukoku> (Accessed: June 10, 2026).

³An example of an ad delivery platform’s guideline: 2. Cosmeceutical Products (Quasi-Drugs) and Cosmetics in Yahoo! Japan Ads Editorial Guidelines: https://ads-help.yahoo-net.jp/s/article/H000044910?language=en_US (Accessed: June 10th, 2026).

Ad Text	Compliance Category
敏感肌に寄り添う、うるおいケア (<i>Moisturizing care tailored to sensitive skin</i>)	✓ Valid
肌の奥から、うるおいで満たす。 (<i>Fills the skin with moisture deep within.</i>)	Overclaiming Effects
ゆらぎ肌ケアに選ばれた美白ライン (<i>Brightening line trusted for sensitive skins.</i>)	Insufficient Information
続けるほど透明感が増す美白美容液 (<i>Brightening serum enhancing radiance.</i>)	Guaranteed Effects
敏感肌も安心の優しさ (<i>Gentle care for sensitive skin, with peace of mind.</i>)	Guaranteed Safety
内側からも外側からも、美肌ケア (<i>Beautiful skin care, inside and out.</i>)	Others

Table 1: Examples from AdCompliance for each compliance category. Since the original advertisements are written in Japanese, English glosses are provided. Underlined text denotes the compliance violation span.

Furthermore, we provide our internal compliance criteria through a taxonomy of rejection reasons to facilitate more fine-grained analysis and support future research and industrial applications. Although these criteria are derived from the Japanese market, the proposed categories are designed to serve as a broadly applicable standard across regions. Nevertheless, additional categories may be required to accommodate local cultural or legal contexts. Therefore, we regard our taxonomy as a basic set of compliance categories that can be extended as needed. As shown in Table 1, we define five basic compliance categories as follows:

Overclaiming Effects Exaggerated claims about product effectiveness, suggestion effects beyond the actual area of application (e.g., “deep within”).

Insufficient Information Claims or phrases that lack supporting evidence or rely on vague achievements or endorsements (e.g., “trusted for”).

Guaranteed Effects Claims that guarantee effectiveness through definitive expressions despite the inherently uncertain nature of the outcome.

Guaranteed Safety Claims that guarantee safety through definitive expressions, implying complete safety or freedom from risk (e.g., “peace of mind”).

Others Other inappropriate expressions that do not fall into the above categories, e.g., issues arising from linguistic nuance or cultural interpretations.

3.2 Dataset Construction

We introduce AdCompliance, an expert-annotated dataset for ad compliance verification. Dataset

Label	Compliance Category	#Example
Valid		6,097
Invalid		1,655
	Overclaiming Effects	520
	Insufficient Information	382
	Guaranteed Effects	109
	Guaranteed Safety	19
	Others	318
	Unspecified	307
Total	–	7,752

Table 2: Number of examples for each label in AdCompliance and the breakdown of compliance categories among invalid ads. Note that some ads are labeled only as *invalid* and categorized as *Unspecified* because the dataset was curated from a recent real AdOps pipeline where rejection reasons had not yet been annotated.

statistics are summarized in Table 2, and representative examples are shown in Table 1. AdCompliance is constructed from real pre-submission ad creatives collected from an AdOps pipeline. Each Japanese advertisement was reviewed as part of the normal AdOps workflow by several in-house experts responsible for compliance verification. Each ad text is assigned a binary valid/invalid label.

In addition, invalid instances are annotated with a diagnostic compliance category and span-level annotations of the rejection reason. These were originally introduced to facilitate agreement among experts. Furthermore, the in-house reviewers regularly participate in consensus-based discussions and feedback sessions to standardize annotation decisions, ensuring comparable results regardless of which expert reviews a particular advertisement. Consequently, although each example is annotated by a single expert in practice, the annotations are aligned with shared organizational standards. Therefore, the labels were effectively double-checked through the consensus process, and no disagreements remained after this process. Furthermore, the dataset construction included a deduplication process, and no annotation inconsistencies were found among duplicate examples.

Note that AdCompliance does not explicitly define fixed training, validation, and test splits, since the dataset is intended primarily for practical industrial use rather than leaderboard evaluation. Instead, we recommend evaluation protocols such as cross-validation, which better reflect practical industry scenarios and provide a more robust assessment across different data partitions. Moreover, as illustrated by the examples in Table 1, the adver-

tisements are written as generic promotional texts without identifiable campaign information. During dataset construction, we also carefully removed duplicate and near-duplicate examples. Therefore, standard split strategies such as stratified splitting can provide sufficiently randomized and representative data partitions.

4 Experimental Settings

Model. We employ Japanese pre-trained encoder models based on ModernBERT (Warner et al., 2025). Specifically, we use two model variants: SB Intuitions ModernBERT (Tsukagoshi et al., 2025) (30M, 70M, 130M, and 310M) and LLM-jp ModernBERT (Sugiura et al., 2025). These models were selected with practical deployment considerations in mind, reflecting the system requirements of integrating into real-world AdOps workflows. We obtained these models from the Hugging Face Transformers library (Wolf et al., 2020).

Training. We evaluate our models using stratified 5-fold cross-validation. For each fold, the training portion is further split into training and validation sets with a ratio of 8:2. We fine-tune the models on AdCompliance for binary valid/invalid classification using joint learning with sequence classification and supervised contrastive learning (Khosla et al., 2021), where the supervised contrastive loss is combined with the classification loss using a mixing weight of 0.2. As shown in Table 2, the label distribution is imbalanced. Therefore, we apply naive random oversampling to the training set per class. We use a learning rate of 2×10^{-4} , a batch size of 256, and early stopping with a patience of 3, terminating training when the validation performance does not improve for three consecutive evaluations. Unless otherwise specified, all implementations are based on the Hugging Face Transformers library (Wolf et al., 2020) and use the default settings of the Trainer API. Training was conducted on a single NVIDIA RTX 3090 GPU.

LLM Comparison. To further investigate whether recent decoder-based models (LLMs) can be employed in the AdOps workflow, we evaluate two Japanese LLMs: CyberAgentLM2-7B-Chat and ELYZA-japanese-Llama-2-7b. Figure 3 shows the prompt used for evaluation, and we apply greedy decoding. We evaluate both zero-shot and few-shot settings. In the few-shot setting, we randomly sample two valid and two invalid

Prompt for LLM Evaluation

以下の基準に基づき、入力された文が日本国内の広告として法的に許容されるかどうかを判定してください。 *[Based on the criteria below, determine whether the input sentence is legally permissible as an advertisement in Japan.]*

判定基準 [Evaluation Criteria]

薬機法（医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律）や景品表示法（不当景品類及び不当表示防止法）を含む、日本の広告関連法規に準拠している、かつ誤認を生じさせない表現である場合は「可」と出力してください。 *[If the sentence complies with Japanese laws and regulations related to advertising, including the Pharmaceuticals and Medical Devices Act (Act on Securing Quality, Efficacy and Safety of Products Including Pharmaceuticals and Medical Devices) and the Act against Unjustifiable Premiums and Misleading Representations, and does not mislead the public, please output "valid."]*

これらの法規のいずれかに違反している場合は「不可」と出力してください。 *[If the sentence violates any of these laws or regulations, please output "invalid."]*

出力形式 [Output Format]

「可」または「不可」のいずれか一方のみで回答してください。回答以外の文字列は絶対に出力しないでください。 *[Respond with only either "valid" or "invalid." Under no circumstances should you output any text other than your response.]*

判定例 [Evaluation Examples]

{few_shot_examples}

判定対象 [Evaluation Target]

入力文：{ad_text} *[Input sentence: {ad_text}]*

出力：[Output:]

Figure 3: Prompt used for LLM inference. Red text is used only in the few-shot setting. English translations are provided as glosses for readability.

examples. These LLM settings, including model selection and prompt, follow Zhang et al. (2025).

Metrics. We assess the results using precision, recall, F1 score, balanced accuracy (Acc.), Matthews correlation coefficient (MCC), and PR-AUC. We report the mean and standard deviation of each metric across the cross-validation folds.

5 Results and Discussions

Main Results. Table 3 shows the overall results. The ModernBERT models consistently achieve high performance on the ad compliance verification task. Moreover, most models obtain an MCC greater than 0.6, indicating a substantial correlation

Model ID	Params.	Precision	Recall	F1 Score	Acc.	MCC	PR-AUC
Chance Rate (All are valid)	–	.787 ± .000	1.00 ± .000	.880 ± .000	.500 ± .000	.000 ± .000	.893 ± .000
<i>Encoder-based Models (ModernBERT)</i>							
sbintuitions/modernbert-ja-30m	37M	.915 ± .004	.944 ± .008	.929 ± .004	.810 ± .008	.649 ± .015	.951 ± .002
sbintuitions/modernbert-ja-70m	70M	.887 ± .050	.962 ± .022	.922 ± .021	.745 ± .123	.526 ± .264	.939 ± .023
sbintuitions/modernbert-ja-130m	132M	.908 ± .008	.953 ± .014	.930 ± .006	.799 ± .016	.646 ± .027	.949 ± .004
sbintuitions/modernbert-ja-310m	315M	.912 ± .010	.948 ± .005	.930 ± .004	.806 ± .019	.649 ± .024	.951 ± .004
llm-jp/llm-jp-modernbert-base	187M	.907 ± .011	.951 ± .015	.928 ± .010	.795 ± .025	.636 ± .048	.948 ± .006
<i>Decoder-based Models (Zero-shot)</i>							
cyberagent/calml2-7b-chat	7B	.785 ± .001	.989 ± .003	.875 ± .001	.496 ± .002	-.034 ± .014	.891 ± .001
elyza/ELYZA-japanese-Llama-2	7B	.787 ± .000	1.000 ± .000	.880 ± .000	.500 ± .000	.000 ± .000	.893 ± .000
<i>Decoder-based Models (Few-shot)</i>							
cyberagent/calml2-7b-chat	7B	.816 ± .015	.337 ± .212	.438 ± .201	.523 ± .008	.044 ± .014	.837 ± .019
elyza/ELYZA-japanese-Llama-2	7B	.773 ± .018	.255 ± .128	.365 ± .155	.491 ± .017	-.017 ± .029	.807 ± .018

Table 3: Prediction performance on the compliance verification task. Since cross-validation is employed, we report the mean and standard deviation over the five folds. Model IDs correspond to their Hugging Face model IDs.

between predictions and ground-truth labels and suggesting that the proposed approach is practical for industrial deployment. In addition, the small standard deviations across the five folds demonstrate the stability and robustness of the proposed approach. Furthermore, performance remains almost consistent across different model sizes, indicating that even lightweight models can be effectively integrated into practical AdOps pipelines for automated compliance verification. In contrast, the LLM-based approaches substantially underperform the encoder-based models, with performance close to the chance level. These results indicate that naively applying recent LLMs is not a practical solution for this task and that simply adopting LLMs does not necessarily provide an effective solution. Considering their model size, unstable performance, and inferior results, there is little motivation to introduce these LLMs into the AdOps workflow for compliance verification. Henceforth, we primarily analyze and discuss LLM-jp ModernBERT as a representative model, since the other ModernBERT models exhibit similar trends.

Confusion Matrix. Figure 4 shows the confusion matrix of LLM-jp ModernBERT. The results indicate that valid advertisements are classified with very few false negatives, meaning that the system rarely rejects compliant ads. In contrast, approximately 70% of invalid ads are successfully identified and filtered. From a practical AdOps perspective, this behavior is desirable. Because multiple ad creatives are typically available and user feedback or personalization can further optimize delivery, it is preferable to avoid unnecessarily rejecting valid ads. Moreover, the primary purpose of this task is

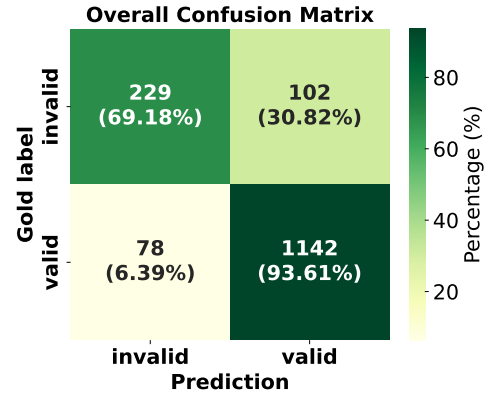


Figure 4: Confusion matrix of LLM-jp ModernBERT. Since the results were stable across the five folds, we report only one fold for analysis. Percentages indicate normalized with respect to the gold labels and therefore correspond to recall for each class.

to assist human experts rather than replace them, as they remain responsible for additional verification tasks and final approval. Therefore, automatically filtering around 70% of non-compliant ads can substantially reduce the manual workload while allowing experts to concentrate on borderline cases. We estimate that it can reduce overall manual verification effort by approximately 15–20%, corresponding to roughly two hours per day under standard operating conditions. This conservative filtering strategy aligns well with the intended role of the system in real-world AdOps.

Confidence. Figure 5 shows box plots of the prediction confidence for each prediction outcome. The results show that correctly classified instances are generally assigned high confidence scores, whereas misclassified instances tend to have rela-

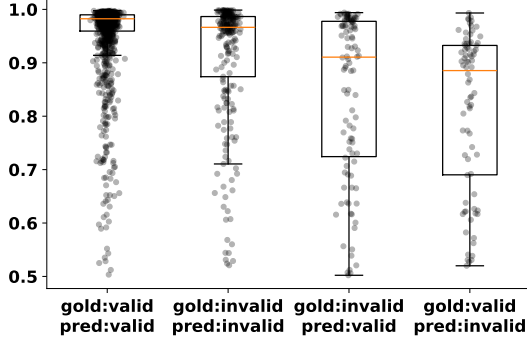


Figure 5: Box plots of prediction confidence for each prediction outcome of LLM-jp ModernBERT.

Ad text	pred.	conf.
美容オイルでハリ・ツヤ・透明感 (A beauty oil for firmness, glow, and radiance.)	invalid	0.980
ダイエット中でも安心のカロリーゼロ (Zero calories, so you can enjoy it even while dieting.)	invalid	0.558
夏の終わりに、紫外線対策を忘れずに (Don't forget UV protection, even at the end of summer.)	valid	0.994
不安な夜を乗り越え、すっきり目覚める (Get through anxious nights and wake up refreshed.)	valid	0.628

Table 4: Examples of misclassified ads by LLM-jp ModernBERT. “pred.” denotes the model prediction, and “conf.” denotes the corresponding prediction confidence.

tively lower confidence. This observation suggests that the confidence score can serve as an effective indicator of prediction reliability. For example, by adjusting the confidence threshold, the system can be configured to conduct more conservatively by forwarding low-confidence cases to human reviewers, or more aggressively by automatically accepting high-confidence predictions, enabling flexible verification strategies depending on requirements.

Qualitative Analysis. Table 4 presents misclassified examples with both high and low confidence scores for each error type. We observe that false negative cases contain definitive expressions. In contrast, false positive errors tend to contain long spans associated with compliance violations. We also observe that some misclassifications are assigned relatively high confidence scores, indicating that confidence alone is insufficient to eliminate all prediction errors. Therefore, in practical deployments, a fully automated decision-making system may be less appropriate than an assistant-style workflow in which the model serves as a pre-screening tool and human experts retain responsibility for final verification and approval.

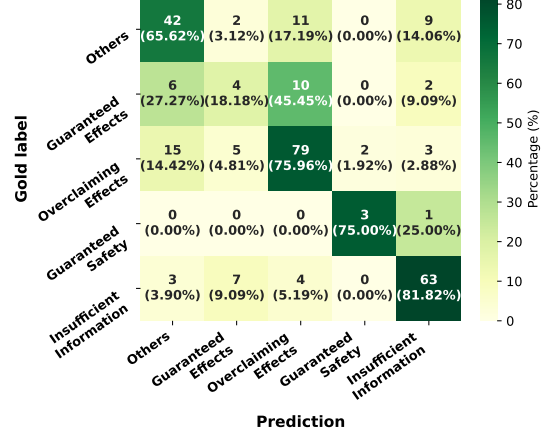


Figure 6: Confusion matrix for compliance category classification by LLM-jp ModernBERT.

Compliance Category Classification. Finally, we investigate whether the model can distinguish compliance categories. Using the same training setup described in Section 4, we extract only instances annotated with one of the five compliance categories and train a five-class classifier. Figure 6 presents the resulting confusion matrix for a fold. This result shows that the model can generally distinguish the compliance categories, even the *Others* class. However, the confusion matrix reveals misclassifications, such as between *Guaranteed Effects* and *Overclaiming Effects*, indicating that some categories remain difficult to reliably separate. In practice, this suggests that the system is better suited to binary valid/invalid verification than to precise compliance categorization, which is also more closely aligned with the requirements of AdOps workflows.

6 Conclusion

In this study, we proposed the ad compliance verification task to improve the efficiency of real-world AdOps workflows and introduced AdCompliance, a dataset for developing and evaluating compliance verification systems. We also developed Japanese ModernBERT-based compliance verification models. Our experiments demonstrated practical classification performance, showing that non-compliant advertisements can be effectively filtered while maintaining a low false positive rate. These findings enabled us to integrate the proposed verification system into the AdOps workflow. As a result, human experts can focus primarily on borderline cases and final decision-making, thereby supporting more efficient and scalable operations.

Limitations

Language. AdCompliance was constructed primarily from Japanese advertisements. Therefore, our experiments and findings are limited to the Japanese language and advertising practices in Japan. However, the contribution of this work is not tied to a particular language, but rather to the formulation of compliance verification as a practical AdOps task that has received relatively little attention in previous advertising research. To promote broader applicability, we designed the basic taxonomy in Section 3.1 to capture general compliance issues such as overclaiming and guaranteed effects, while including an *Others* category for language- or culture-specific cases. Nevertheless, this paper does not empirically evaluate the transferability of the proposed task and taxonomy to other languages or cultural contexts, such as Chinese, Korean, or regions with substantially different advertising regulations. Therefore, we leave multilingual and cross-cultural validation for future work. That said, given its focus on practical industrial use, we believe that introducing compliance verification as a new AdOps task and providing real-world insights into this problem constitute meaningful contributions.

Model. Our purpose is to integrate the proposed system into real-world AdOps workflows. Therefore, we primarily selected ModernBERT with practical deployment constraints and production environments in mind. Moreover, as shown in Section 5, even relatively small models achieved good performance on the compliance verification task. Moreover, the primary contribution of this work is the formulation of the task and its practical application rather than benchmarking model capability. Therefore, we did not conduct extensive comparisons with models or other architectures. Consequently, based on the discussions above, although we expect similar trends to hold for alternative model families and smaller models, their actual performance remains unverified.

Applicability. Finally, one possible limitation is that this paper demonstrates the practical feasibility of the proposed task, but does not extensively explore its broader research potential beyond the immediate industrial use case. As an industry-focused paper, our primary contribution is the introduction of a novel application and use case in AdOps. Nevertheless, the proposed task has several promising directions for future research, e.g., multilingual

compliance verification, extension to multimodal advertising content such as landing pages, banner advertisements, and commercial videos, and the use of compliance verification models as reward signals for ad generation systems. While these possibilities suggest that the task has broader applicability than the setting considered in this paper, we do not empirically investigate or validate such extensions. Exploring these directions remains future work. Despite this limitation, we believe that establishing compliance verification as a fundamental AdOps task provides a useful foundation for future academic research and industrial applications.

Ethical Considerations

License Agreements. AdCompliance was constructed from data collected through our AdOps workflow. We obtained permission from the corresponding advertisers to use the ad texts for research purposes. In addition, we obtained permission to include representative examples in this paper, provided that no information capable of directly identifying specific products or advertisers is disclosed. The annotation results are owned by our organization because the expert annotations were produced as part of the normal business workflow. Therefore, we do not identify any licensing or ownership issues regarding the use of the dataset in this study. Regarding the models, SB Intuitions ModernBERT is distributed under the MIT License, while LLM-jp ModernBERT and the Hugging Face Transformers library are distributed under the Apache License 2.0. We verified compliance with all applicable licenses. Furthermore, any released materials will be accompanied by appropriate licenses.

Harmful Contents. We did not intentionally generate, promote, or disseminate harmful content in any part of our experiments. Although false positives and false negatives may occur, the final decision is always made by human experts. Our system is designed solely to assist human reviewers rather than replace them. Thus, classification errors do not directly affect customers. We confirm that this research complies with the ACL Code of Ethics⁴.

AI Tools. We used AI tools solely for grammatical correction and translation assistance; all research ideas, experimental design, and original content were developed by the authors. The authors take full responsibility for the content of this work.

⁴<https://ethics.aclweb.org/>

References

- Yanda Chen, Zihui Ren, Qixiang Gao, Jiale Chen, Si Chen, Xubin Li, Tiezheng Ge, and Bo Zheng. 2025. [Ctr-driven ad text generation via online feedback preference optimization](#). Preprint, arXiv:2507.20227.
- Go Inoue, Akihiko Kato, Masato Mita, Ukyo Honda, and Peinan Zhang. 2024. [CAMERA³: An evaluation dataset for controllable ad text generation in Japanese](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2702–2707, Torino, Italia. ELRA and ICCL.
- Hidetaka Kamigaito, Soichiro Murakami, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. 2024. [Generating attractive ad text by facilitating the reuse of landing page expressions](#). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 597–608, Tokyo, Japan. Association for Computational Linguistics.
- Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. 2021. [An empirical study of generating texts for search engine advertising](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 255–262, Online. Association for Computational Linguistics.
- Akihiko Kato, Masato Mita, Soichiro Murakami, Ukyo Honda, Sho Hoshino, and Peinan Zhang. 2024. [Faith-camera: Construction of a faithful dataset for ad text generation](#). Preprint, arXiv:2410.03839.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. [Supervised contrastive learning](#). Preprint, arXiv:2004.11362.
- Zeyang Lei, Chao Zhang, Xinchao Xu, Wenquan Wu, Zheng-yu Niu, Hua Wu, Haifeng Wang, Yi Yang, and Shuanglong Li. 2022. [PLATO-ad: A unified advertisement text generation framework with multi-task prompt learning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 512–520, Abu Dhabi, UAE. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Zebo Liu, Kehan Li, Xu Tan, and Jiming Chen. 2020. [IAD: A Benchmark Dataset and a New Method for Illegal Advertising Classification](#). In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, pages 2085–2092, Santiago de Compostela, Spain. IOS Press.
- Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. 2024. [Striking gold in advertising: Standardization and exploration of ad text generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 955–972, Bangkok, Thailand. Association for Computational Linguistics.
- Soichiro Murakami, Sho Hoshino, and Peinan Zhang. 2023. [Natural language generation for advertising: A survey](#). Preprint, arXiv:2306.12719.
- Soichiro Murakami, Peinan Zhang, Sho Hoshino, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2022. [Aspect-based analysis of advertising appeals for search engine advertising](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 69–78, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.
- Soichiro Murakami, Peinan Zhang, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2025a. [AdParaphrase: Paraphrase dataset for analyzing linguistic features toward generating attractive ad texts](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1426–1439, Albuquerque, New Mexico. Association for Computational Linguistics.
- Soichiro Murakami, Peinan Zhang, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2025b. [AdParaphrase v2.0: Generating attractive ad texts using a preference-annotated paraphrase dataset](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15212–15230, Vienna, Austria. Association for Computational Linguistics.
- Hiroto Otake, Peinan Zhang, Yusuke Sakai, Masato Mita, Hiroki Ouchi, and Taro Watanabe. 2025. [BannerBench: Benchmarking vision language models for multi-ad selection with human preferences](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24145–24159, Suzhou, China. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Issa Sugiura, Kouta Nakayama, and Yusuke Oda. 2025. [llm-jp-modernbert: A modernbert model trained on a large-scale japanese corpus with long context length](#). Preprint, arXiv:2504.15544.
- Yi-jie Tang and Hsin-Hsi Chen. 2014. [FAdR: A system for recognizing false online advertisements](#). In *Proceedings of 52nd Annual Meeting of the*

- Association for Computational Linguistics: System Demonstrations, pages 103–108, Baltimore, Maryland. Association for Computational Linguistics.
- Hayato Tsukagoshi, Shengzhe Li, Akihiko Fukuchi, and Tomohide Shibata. 2025. [ModernBERT-Ja](https://huggingface.co/collections/sbintuitions/modernbert-ja-67b68fe891132877cf67aa0a). <https://huggingface.co/collections/sbintuitions/modernbert-ja-67b68fe891132877cf67aa0a>.
- Chang Wang, Siyu Yan, Depeng Yuan, Yuqi Chen, Yanhua Huang, Yuanhang Zheng, Shuhao Li, Yinqi Zhang, Kedi Chen, Mingrui Zhu, and Ruiwen Xu. 2025a. [Beyond quality: Unlocking diversity in ad headline generation with large language models](#). Preprint, arXiv:2508.18739.
- Heng Wang, Yotaro Shimose, and Shingo Takamatsu. 2025b. [BannerAgency: Advertising banner design with multimodal LLM agents](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4304–4329, Suzhou, China. Association for Computational Linguistics.
- Zhao Wang, Bowen Chen, Yotaro Shimose, Sota Moriyama, Heng Wang, and Shingo Takamatsu. 2025c. [Mirror in the model: Ad banner image generation via reflective multi-LLM and multi-modal agents](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 246–266, Suzhou (China). Association for Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Yexing Xu, Wei Feng, Shen Zhang, Haohan Wang, Yuxin Qin, Yaoyu Li, Ao Ma, Yuhao Luo, Lu Wang, Xudong Ren, Haoran Wang, Run Ling, Zheng Zhang, Jingjing Lv, Junjie Shen, Ching Law, Longguang Wang, and Yulan Guo. 2026. [Design your ad: Personalized advertising image and text generation with unified autoregressive models](#). Preprint, arXiv:2605.12138.
- Peinan Zhang, Yusuke Sakai, Masato Mita, Hiroki Ouchi, and Taro Watanabe. 2025. [AdTEC: A unified benchmark for evaluating text quality in search engine advertising](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7672–7691, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yingxue Zhou, Tan Zhu, Tao Zeng, Zigeng Wang, and Wei Shen. 2025. [Learning from LLM agents: In-context generative models for text casing in E-commerce ads](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1122–1133, Suzhou (China). Association for Computational Linguistics.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. [Txygen: A benchmarking platform for text generation models](#). Preprint, arXiv:1802.01886.