AdTEC: A Unified Benchmark for Evaluating Text Quality in Search Engine Advertising

Anonymous ACL submission

Abstract

With the increase in the more fluent ad texts automatically created by natural language generation technology, it is in the high demand to verify the quality of these creatives in a real-world setting. We propose AdTEC, the first public benchmark to evaluate ad texts in multiple aspects from the perspective of practical advertising operations. Our contributions are: (i) Defining five tasks for evaluating the quality of ad texts, as well as building a Japanese dataset based on the practical operational experiences of advertising agencies, which are typically kept in-house. (ii) Validating the performance of existing pre-trained language models (PLMs) and human evaluators on the dataset. (iii) Analyzing the characteristics and providing challenges of the benchmark. The results show that while PLMs have already reached the practical usage level in several tasks, human still outperforms in certain domains, implying that there is significant room for improvement in such area.

1 Introduction

Online advertising, especially sponsored search advertising, is a dominant sector for vendors to promote their products, and the market size is estimated to grow by billions of dollars over the next few years (Murakami et al., 2023). To meet the increasing demands of advertising operations (AdOps), such as creating ad texts from product information (Step 2 in Figure 1), the remarkable success of natural language generation (NLG) by pre-trained language models (PLMs) (Dong et al., 2021; Vaswani et al., 2017; Brown et al., 2020; Touvron et al., 2023a) has given a boost to applications in practice (Hughes et al., 2019; Kamigaito et al., 2021; Golobokov et al., 2022), making advertising a huge industrial use case for NLP.

In this paper, we focus on evaluating the quality of ad texts generated by such models. We refer to this process as *ad text evaluation*, as depicted in Step 3 of Figure 1. Ad text evaluation is crucial because low-quality ad texts can be a detriment to advertisers through a lack of fluency, inappropriate appeals, and misleading representations. Because it is expensive and non-scalable to ask humans for verifying the quality of each text in high-volume domains such as sponsored search advertising, it is in the high demand for developing automatic quality estimators for ad texts. The quality has multiple dimensions, such as appropriate wording, effective appeals, consistency between ad text and product information, and high predicted performance. Although these dimensions should be included for evaluating automatic quality estimators, there is no such benchmark. Consequently, the bottleneck lies in verifying the quality of ad texts despite the ability of generating numerous creatives automatically, which hinders the scalability of delivery volume. Thus, we aim to construct a benchmark for evaluating the quality of ad texts.

The primary challenge in constructing an ad text evaluation benchmark is the absence of a clear definition for the tasks (Murakami et al., 2022; Mita et al., 2024). The lack of domain knowledge in AdOps complicates the understanding of high-quality ad text standards and the accurate definition of tasks. However, the AdOps workflow is complex, relies on various platforms and formats, and encompasses multiple methodologies and metrics. In addition, only a few companies possess the expertise to operate online advertising at scale. Owing to legal and contractual obligations, advertising workflows and data are predominantly managed in-house, leading to a lack of publicly available datasets. This scarcity makes it challenging to systematically reproduce and validate diverse methodologies in academia. Consequently, research remains less active in the advertising field, potentially overlooking issues

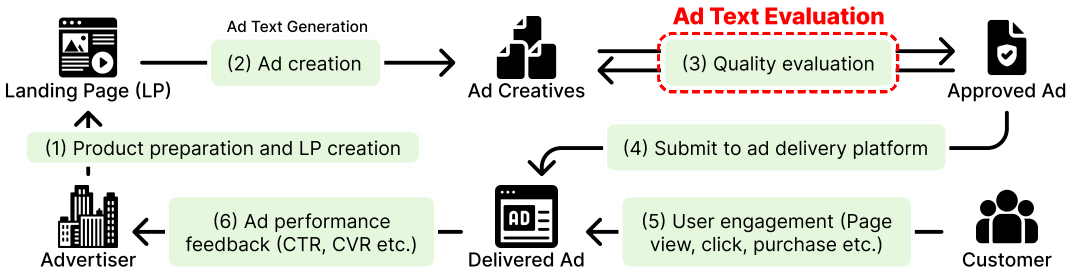

Figure 1: Generalized AdOps workflow described in §2: (1) The advertiser creates an LP to promote a product. (2) Based on the product information in the LP and target customers, text and graphics are designed by creators. (3) The creatives are evaluated based on fluency, attractiveness, regulations, legality and other factors. (4) Once the creatives pass the quality evaluation, they are submitted to a delivery platform. (5) Customers take actions on the displayed ads, such as page views, clicks, and purchases. (6) Based on the customer engagement, ad performance is reported back to the advertiser and returned to Step 1 to improve the quality of the LP and ads.

and delaying the application and development of cutting-edge technology.

To address these challenges, we propose **ADTEC**, the first public benchmark that defines and unifies tasks based on generalized AdOps workflows. Our major contributions are as follows:

The First Public Dataset on Ad Text Evaluation. We organized real-world advertising workflow and carefully designed five tasks closely match this workflow to evaluate the quality of ad texts. This was based on the practical operational experiences of advertising agencies, using Japanese data as a case study. We are releasing this dataset at <URL>, marking it as the first publicly available dataset for ad text evaluation, as such data is typically kept in-house and difficult to obtain.

Benchmark Experiments. We validated the performance of existing PLMs, such as BERT, RoBERTa, and large language models (LLMs), as well as human evaluators on our proposed benchmark.

Dataset Analysis. We analyzed the characteristics and identified potential issues of the dataset through experiments, demonstrating that our benchmark is challenging, and highlighting potential areas for improvement and future research.

2 Understanding Sponsored Search Advertising and its Workflow

Sponsored search advertising displays titles and descriptions of ads that are relevant to keywords, which users enter into search engines, appearing as part of the search results. As illustrates in Figure 2, when a user clicks ad's URL, they are directed

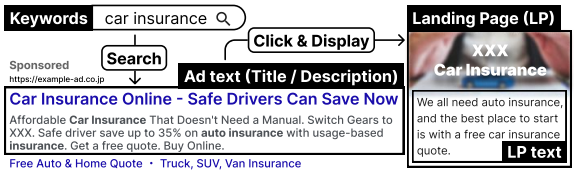

Figure 2: Overview of sponsored search ad.

to a web page known as a landing page (LP). LP contains texts and images related to the advertiser's products, allowing the user to take actions, such as making a purchase.

Operating such ads requires high expertise, owing to the variety of delivery platforms, properties and formats. To gain insights, we interviewed two types of experts familiar with AdOps; those overseeing AdOps departments at ad agencies and those directly involved in on-site operations. Through these interviews, we explored the intricacies of AdOps, generalized the workflow, and delineated six steps, as illustrated in Figure 1. In this work, we focus mainly on Step 3 of quality evaluation.

3 Building ADTEC

Our goal is to develop a benchmark that captures multi-dimensional aspects of quality that are relevant to practical scenarios and carefully curated to provide value for both real-world applications and various research purposes.

3.1 Task Design

We followed the workflow outlined in §2, and asked the same experts from the interview to identify crucial parts within Step 3 in Figure 1. Derived from their insights, we generalized and defined tasks that adhere to the principle of evaluating ad texts directly or indirectly. *Direct* evaluation tasks

Ad text	LP text	AD ACCEPT. / CONSIST.
Condominium Sales / Free Assessment Now	Let XX assess your single-family home!	`acceptable` / `inconsistent`
Engineer's career / Engineer's Job	Find jobs at website XX!	`unacceptable` / `consistent`

Table 1: Examples of AD ACCEPTABILITYand AD CONSISTENCY task. Note that LP texts are only used in the AD CONSISTENCY task.

are used to test texts with strict criteria, such as a binary pass/fail outcome or a numerical score to quantify the text quality. These tasks will serve as a checklist to ensure that minimum delivery standards are met. *Indirect* evaluation tasks are used to assist human evaluators in reviewing or refining texts or serve as a bridge to connect the downstream tasks. Based on the above principles, we designed five tasks: three direct evaluation tasks (AD ACCEPTABILITY, AD CONSISTENCY, and AD PERFORMANCE ESTIMATION) and two indirect evaluation tasks (A^3 RECOGNITION and AD SIMILARITY).[1]

AD ACCEPTABILITY. As most ad delivery platforms impose text length restrictions, minor grammatical errors are tolerated to enhance the readability and engage readers within limited space. However, excessive compression can mislead readers, and such poor-quality ads should be detected before delivery to avoid negative impacts on the advertiser. To assess this, we defined the task of AD ACCEPTABILITY, which predicts the acceptance of overall quality with binary labels: `acceptable`/`unacceptable`. Unacceptable ad phenomena include "collapsing symbols," "unnatural repetition," and "incomprehensible meaning" based on expert feedback. This differs from the general concept of linguistic acceptability, which checks for grammatical correctness, such as CoLA (Warstadt and Bowman, 2019). Examples of AD ACCEPTABILITY are shown in Table 1. The ad text "*Engineer's career / Engineer's Jobs*" is *unacceptable* because the meaning is duplicated.

AD CONSISTENCY. Verifying consistency between the ad text and LP content is crucial. If a feature or price mentioned in the ad text is not referenced in the corresponding LP, it may violate the Law for the Prevention of Unjustified Extra or Unexpected Benefit and Misleading Representation, resulting in damages to the advertiser. However, these inconsistencies are hard to detect as some factual expressions do not appear in LPs. For example, the term "*official*," which is often used in ad text but rarely appears in LP content. To assess this, we defined the AD CONSISTENCY task, which predicts the consistency between LP content and ad text with binary labels: `consistent`/`inconsistent`. Examples of AD CONSISTENCY are shown in Table 1. The first line is labeled inconsistent because the LP refers to a "*single-family home*," while the ad text mentions a "*condominium*."

AD PERFORMANCE ESTIMATION. The most straightforward way to measure ad quality is to publish them online and let end customers evaluate them. However, delivering all ads without alterations is impractical, as low-quality ads can negatively impact advertisers. Therefore, prior studies investigated offline methods to measure ad text quality by simulating customer behavior, such as click-through rate (CTR), based on past delivery history. These methods are currently standard practices in many organizations. Inspired by their works (Gharibshah et al., 2020; Niu and Hou, 2020; Yang and Zhai, 2022), we adopted AD PERFORMANCE ESTIMATION task to estimate a quality score in the range of [0, 100] from ad texts, keywords, and industry, as shown in Table 2. The score simulates customer behavior based on past delivery history, and is non-linearly transformed to maintain the original label distribution for contractual reasons.

A^3 RECOGNITION. One of the most crucial factors in advertising is the *aspect of advertising appeals* (A^3). At its core, advertising aims to connect advertisers with readers, and A^3 serves as a bridge between them. For example, an ad emphasizing *low cost* may resonate with price-conscious readers, while one focusing on *high performance* may not. Thus, recognizing appealing expressions in advertising and using appropriate A^3 can enhance downstream tasks, such as CTR prediction (Murakami et al., 2022). In the A^3 RECOGNITION task, we follow a previous study

[1]Note that all examples in this paper are translated from Japanese into English for the presentation brevity.

Field	Example value
Title 1	[No.1] Card loan comparison site
Title 2	A must-see for those who in a hurry!
Title 3	Instant Loan Secure Card Loan
Desc. 1	The best place to get a card loan without telling anyone. You only need a driver's license to apply
Desc. 2	Convenient to use ATMs at convenience stores. Convenient and quick loans are available if you apply before 10:00 p.m.
Keyword	card loan
Industry	finance
Score	82.3

Table 2: Examples of AD PERFORMANCE ESTIMATION task. Desc. represent Description. Score is the label and others are inputs.

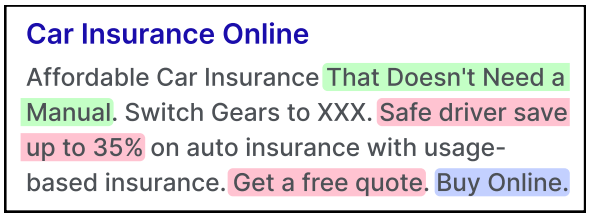

Figure 3: Example of A^3 RECOGNITION task. The highlight area indicates the A^3 : *Features*, *Special deals*, and *User-friendliness*.

	Sentences	Score
S1	Suppon Black Vinegar with Luxury Ceramide	5.00
S2	Suppon Black Vinegar and Luxury Ceramide	
S1	Find a gift that fits your budget	2.33
S2	Save up to 40% on discounted products	

Table 3: Examples of AD SIMILARITY task. S1 and S2 represent the paired Sentences 1 and 2, respectively.

(Murakami et al., 2022), which predicts all relevant A^3 labels in a given ad text. Figure 3 shows an example of ad texts and the corresponding A^3s. All labels and distributions can be found in Table 11.

AD SIMILARITY. Repeatedly showing same ads to readers leads to *ad fatigue* (Abrams and Vee, 2007), where readers become bored and ad performance declines. Therefore, it is essential to avoid displaying the same ads for extended periods, and regularly replace them with different ones. However, the transition from old to new ads must be carefully managed because we need to maintain the product and its appeal while updating the wording or representations, or we risk disengaging customers who were attracted to the previous ads. Thus, measuring the similarity particularly focusing on this situation is crucial, which enables us to determine whether to replace the ad based on a quantified score.

Building on the aforementioned motivation, we defined the AD SIMILARITY task, which predicts the similarity score for an ad text pair on a scale of [1, 5]. The lower the values, the less similar the pair, and vice versa. Examples are shown in Table 3. The first example pair illustrates high similarity, where the both the product of "*Suppon Black Vinegar*" and the A^3 of *luxury* are identical. Conversely, the second example pair differ in A^3 with *budget* and *discount*, resulting in relatively low similarity.

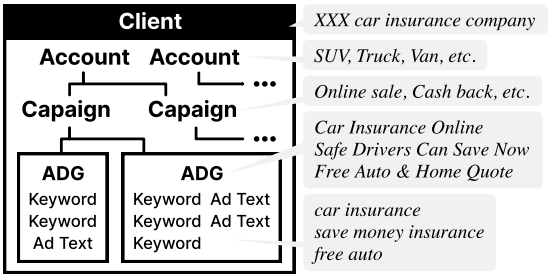

Figure 4: Account structure in ad delivery is hierarchical. A client represents a single company, and the account typically encompasses the commercial products offered by that client. Campaigns are created to promote these commercial products, while ad groups are used to organize keywords and ad texts. At higher levels of the hierarchy, there are more ads and greater variance. Conversely, at lower levels, there are fewer ads, which tend to be more similar.

3.2 Dataset Construction

Data Collection. For the AD ACCEPTBILITY and AD CONSISTENCY task, data was collected from the ad creation phase of the actual AdOps workflow, including both human creators and NLG models' outputs. For the AD PERFORMANCE ESTIMATION and AD SIMILARITY task, we used Japanese sponsored search ads delivered between 2021 and 2022. For the A^3 RECOGNITION task, we used data from Murakami et al. (2022).

Data Pre-processing. In the AD PERFORMANCE ESTIMATION task, we negotiated with our clients to use a subset of the data approved for public release, as the data includes ad performance metrics sensitive to advertisers. Furthermore, we applied a nonlinear transformation to the raw CTR, scaling them to the range of $[0, 100]$ to preserve their

Task	Setup: Input → Label	Metrics
AD ACCEPTABILITY	Classification: Ad text → `acceptable/unacceptable`	Accuracy/F1-score
AD CONSISTENCY	Classification: (Ad text, LP text) → `consistent/inconsistent`	Accuracy/F1-score
AD PERF. EST.	Regression: (Ad texts, Keyword, Industry) → $[0, 100]$	Pearson/Spearman corr.
A^3 RECOGNITION	Classification: Ad text → multi-labels (See Appendix Table 11)	F1-micro/-macro
AD SIMILARITY	Regression: Ad text pair → $[1, 5]$	Pearson/Spearman corr.

Table 4: Task descriptions of ADTEC. AD PERF. EST. and Corr. are abbreviations for AD PERFORMANCE ESTIMATION and correlation coefficient, respectively.

Task	Train	Dev	Test
AD ACCEPTABILITY	13,265	970	980
AD CONSISTENCY	10,635	945	970
AD PERF. EST.	125,087	965	965
A^3 RECOGNITION	1,856	465	410
AD SIMILARITY	4,980	623	629

Table 5: Number of instances for each task in each dataset.

distribution. Additionally, we masked proper nouns (e.g., product names, company names) to prevent identification of advertisers and avoid any potential negative repercussions upon data release. In the AD SIMILARITY task, creating sentence pairs through random sampling is inefficient because most pairs are not similar. Therefore, we utilized the account structure configured during ad delivery, which includes *client*, *account*, *campaign*, *ad group*, and *keyword* information, as depicted in Figure 4. We created pseudo-similar pairs by sampling texts from the same *ad group*, as preliminary results suggested. To balance the label distribution, we created pseudo-dissimilar pairs by ensuring that the two texts belonged to different *clients*. Consequently, we sampled pseudo-similar and pseudo-dissimilar pairs at a ratio of 9:1.

Annotation Workflow. All annotators are native Japanese speakers and experts with vast experience in AdOps. The annotation workflow for all tasks, except AD PERFORMANCE ESTIMATION and A^3 RECOGNITION, is as follows: (1) We first removed duplicate entries and filtered out data in languages other than Japanese. (2) We iteratively revised the annotation guidelines until we achieved a satisfactory level of agreement through pilot annotations on small sampled datasets. (3) Following the guidelines, we conducted a pilot annotation by asking three annotators to annotate the same sampled data used in Step 2. Thereafter, we compared the annotators' results with our own and resolved any inconsistencies to further refine the guidelines. This cycle was repeated at least twice. (4) After completing Steps 1 to 3, we conducted the main annotation on the full test set using the finalized guidelines. The complete guidelines are provided in the Appendix B.

Data Splitting. Despite efforts at deduplication, similar ad expressions can still be easily found, potentially leading to data leakage with a simple random split. Furthermore, assuming the data is used in industry, it is crucial to generalize effectively without overfitting to specific ad expressions. Therefore, for AD ACCEPTABILITY and AD CONSISTENCY, we split the data, considering the ad hierarchical structure in Figure 4, ensuring that *clients* do not overlap across training, development, and testing. For the AD PERFORMANCE ESTIMATION, we used the delivery structure, with the non-overlapping layer as the campaign. For the A^3 RECOGNITION, we used the same split from Murakami et al. (2022). For the AD SIMILARITY, we randomly split the data to maintain label distribution consistency.

Table 5 provides the statistics of our dataset. The full tables of label distribution for each task are provided in Appendix Table 11.

4 Experiment

4.1 Experimental Settings

Table 4 provides a brief overview of each task and the corresponding metrics used to measure task performance. Our tasks are categorized into three setups: binary classification, multi-label classification, and regression. For binary classification, we use accuracy and the F1-score to evaluate the binary labels. In multi-label classification, we follow Murakami et al. (2022) and use the F1-score with both macro and micro settings. For regression, we use Pearson and Spearman correlation coefficients.

We employed two types of evaluators: the PLMs, including both fine-tuned settings with encoder models and zero-/few-shot settings for

Evaluator	AD ACCEPT. Accuracy/F1-score	AD CONSIST. Accuracy/F1-score	AD PERF. EST. Pearson/Spearman	A^3 RECOGNITION F1-micro/-macro	AD SIMILARITY Pearson/Spearman
			Fine-tuned Encoder Models		
Tohoku BERT	0.711/0.688	**0.767**/0.552	**0.480/0.497**	0.774/**0.694**	0.773/0.807
Waseda BERT	0.598/0.637	0.755/0.474	0.445/0.457	0.663/0.517	0.740/0.800
XLM-RoBERTa	0.705/0.690	0.758/0.519	0.453/0.457	**0.778**/0.677	**0.878/0.878**
			Zero-/Few-shot LLMs		
$CALM2_{7b}$	0.520/0.115	0.381/0.472	0.006/0.013	0.154/0.042	0.036/0.036
$ELYZA_{7b}$	0.352/0.520	0.628/0.771	0.003/0.046	0.196/0.044	0.015/-0.004
GPT-4	0.325/0.433	0.583/0.612	0.028/0.073	0.417/0.113	0.776/0.811
Human	**0.732/0.790**	0.703/**0.807**	—	0.564/0.538	0.699/0.765

Table 6: Performance of PLMs and human evaluators on the test set. Underlined indicates the best result for each setting, and **bold** indicates the best result across all methods. (See Appendix Table 12 for the full results.)

LLMs, as well as human evaluators. Please not that the detailed model description, hyperparameters, prompts, and additional information are provided in Appendix C.

Fine-tuning Setting with Encoder Models. We utilized publicly available encoder models as baselines, namely Tohoku BERT, Waseda RoBERTa, and XLM-RoBERTa (Conneau et al., 2019), commonly employed for Japanese NLP tasks, which differ in pre-tokenizer, tokenization unit, and pre-training dataset. All the aforementioned models are of LARGE size.

Zero-/Few-shot Setting with LLMs. We employed $CALM2_{7b}$ and $ELYZA_{7b}$as baselines for open LLMs. $CALM2_{7b}$ and $ELYZA_{7b}$ are based on the Llama 2 (Touvron et al., 2023b) architecture but differ in training methods and data; $CALM2_{7b}$ was trained from scratch, whereas $ELYZA_{7b}$ was continuously trained from the original Llama 2.[2]

Human. We enlisted three human evaluators who are *not* engaging in AdOps to evaluate all tasks except AD PERFORMANCE ESTIMATION. We followed the same procedure for instruction as described in §3.2. Pilot evaluations were carried out twice on randomly sampled 100 instances from the training set for each task before the main run. In the final assessment, a majority vote per instance was conducted for AD ACCEPTABILITY and AD CONSISTENCY tasks, while the average scores of evaluator assessments were reported for other tasks.

[2]To further assess the gap between fine-tuning and zero-/few-shot learning of LLMs, we also fine-tuned an LLM. Due to the additional nature of this experiment, which falls beyond the scope of our study and is constrained by our computational resources, we selected only the best-performing open-source model from the zero-/few-shot settings. The settings and results are provided in Appendix C.

4.2 Results and Discussion

Table 6 provides an overview of the result. For the complete result including base size encoder models and GPT-3.5, please refer to Appendix Table 12.

Fine-tuned Setting with Encoder Models. XLM-RoBERTa and Tohoku BERT achieved the highest or competitive scores in two or more tasks. In addition, the LARGE model outperformed the BASE model in most tasks, suggesting that the increasing parameter size plays a key role in understanding ad expressions.

Zero-/Few-shot Setting with LLMs. GPT-4 achieved high performance across all three tasks, while $ELYZA_{7b}$ performed the best among the LLMs in AD CONSISTENCY. A substantial difference was observed between the open and OpenAI's LLMs in the AD SIMILARITY task. $CALM2_{7b}$ and $ELYZA_{7b}$ scored close to 0, indicating no correlation, whereas OpenAI's models, especially GPT-4, achieved competitive scores of 0.776/0.811. Thus, the open LLMs used in this study struggle with handling the semantic similarity or numerical answers, whereas GPT-4 performs considerably better. In the AD PERFORMANCE ESTIMATION task, all LLMs produced uncorrelated responses close to 0, indicating that accurately predicting ad performance remains a challenge for open LLMs.

Fine-tuned Encoder Models vs. Zero-/Few-shot LLMs. Overall, the fine-tuned models outperformed the LLMs. The difference was substantial, ranging from 0.2 to 0.6, especially

for the AD ACCEPTABILITY, AD PERFORMANCE ESTIMATION, and A^3 RECOGNITION tasks. This can be attributed to the characteristics of the tasks; AD PERFORMANCE ESTIMATION involves predicting numbers in the range $[0, 100]$, and A^3 RECOGNITION requires selecting all suitable labels from more than 20 labels, suggesting that the variety of data features and outputs could not be handled by few-shot alone. Both AD ACCEPTABILITY and AD CONSISTENCY are binary classification tasks, yet the performance gap between fine-tuned models and LLMs is more noticeable in the AD ACCEPTABILITY task, with a difference as large as 0.2 points, compared to the relatively smaller difference observed in the AD CONSISTENCY task. The task with the smallest difference was AD SIMILARITY, with a difference of only 0.06 to 0.10.

PLMs vs. Human. Human evaluators outperform models in AD ACCEPTABILITY and AD CONSISTENCY. In both tasks, models tends to have high accuracy and a low F1-score, while humans exhibit the opposite trend. Particularly in the AD CONSISTENCY task, where the label distribution is unbalanced, even the best model achieves an F1-score of only 0.55, whereas human performance reaches 0.8. This suggests that humans can make better predictions in both precision and recall in an unbalanced data. On the other hand, fine-tuned models outperform humans on the A^3 RECOGNITION and AD SIMILARITY tasks. In the A^3 RECOGNITION task, human evaluators struggled with the diversity of output labels, similar to LLMs. However, the difference between the F1-micro and F1-macro scores is relatively small at 0.03 points for the human result. In contrast, for the fine-tuned PLMs, the difference ranges from 0.08 to 0.20 points. This indicates that human evaluators have a strong ability to generalize and can maintain higher performance even when a label appears infrequently. In the AD SIMILARITY task, in addition to fine-tuned models, GPT-4 outperforms human evaluators. This is the only task where LLM outperforms humans, suggesting that GPT-4 has a high level of alignment with humans in terms of semantic similarity and numerical understanding.

Error Analysis on AD ACCEPTABILITY and AD CONSISTENCY Task. We have conducted a detailed analysis of two tasks, AD ACCEPTABILITY and AD CONSISTENCY, where the model has not yet outperformed human evaluators. We analyzed three types of errors: human incorrect and model correct (IC-C), human correct and model incorrect (C-IC), and both incorrect (IC-IC), as shown in Table 7. In both tasks, the model more often predicted False labels than True labels for both IC-C and C-IC errors.. In contrast, humans provided more True labels in both correct and incorrect cases. This suggests that humans exhibit a relatively higher degree of leniency compared to models, which tend to be overly cautious when making decisions in AD ACCEPTABILITY and AD CONSISTENCY tasks.

Error	GT	H	M	ACCEPT.	CONSIST.
C-IC	T	T	F	30.1%	17.2%
	F	F	T	0.6%	5.5%
IC-C	T	F	T	0.6%	1.9%
	F	T	F	18.0%	10.9%
IC-IC	T	F	F	4.9%	6.2%
	F	T	T	3.9%	6.6%

Table 7: Type-specific error rates for Tohoku BERT (BERT) and XLM-RoBERTa (XLM-R) in the AD ACCEPTABILITY and AD CONSISTENCY task. Ground truth (GT), human (H), and model (M) labels are represented as T (`acceptable` or `consistent`) and F (`unacceptable` or `inconsistent`).

Case Study on the Model's Behavior on AD SIMILARITY Task. We used integrated gradients (Sundararajan et al., 2017) to visualize the attention of Tohoku BERT, aiming for a deeper understanding of the model's behavior, which can potentially enhance its performance. Figure 5 shows examples of low and high errors between predictions and the ground truth in the AD SIMILARITY task. In the first example, the model fails to pay sufficient attention to "*online seminar*," which has a similar meaning to "*webinar*." However, high attention to the date expression "*9/9 (Thursday) at 13:00*" allows the model to predict a score close to the correct answer. However, in the second example, the model did not pay enough attention to "*English conversation coaching*," which has a similar meaning to "*gaining English skills*," resulting in an incorrect prediction of dissimilarity. As we found many such cases apart from the ones mentioned, we believe that the model tends to prioritize surface-level information, especially for entities such as dates, times, and numbers. This may prevent the model from

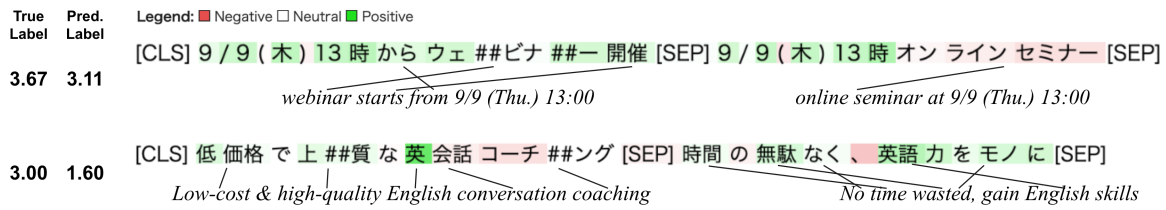

Figure 5: Examples of integrated gradient visualization with Tohoku BERT model's outputs showing the difference in attention for small (top) and large (bottom) gaps between ground truth and predicted labels in AD SIMILARITY task. Red indicates negative influence, while green indicates positive influence in the predictions.

correctly identifying cases that are semantically similar even when the surface-level information differs, and vice versa.

Another instance where we observed shortcomings in the models is in reasoning. For example, the models mistakenly interpret that phrases such as "*half price*" and "*buy 2 get 1 free*" have the same meaning, whereas humans easily comprehend. Additionally, pairs of phrases like "*Available 24 hours a day, 7 days a week*" and "*Closed every Wednesday*," and "*Property located near the station*" and "*20-minute walk from the station*" often confuse the models in understanding. These types of paraphrasing are commonly used in the ad creation process; however, it is challenging to automatically capture them in a natural dataset. Therefore, a specialized dataset focusing on these phenomena is necessary to enable models to capture them accurately.

5 Related Work

Mita et al. (2024) proposed an ad text generation benchmark that aims to evaluate NLG models (Step 2 in Figure 1) with surface- and semantic-level overlap metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020), along with simple heuristics, including keyword insertion rate (Mishra et al., 2020) and text length. However, this approach is inadequate for real-world applications because: (1) it requires reference texts to evaluate generated texts, making it difficult to evaluate new ads that have no references, and (2) simple heuristics cannot guarantee the quality of the ad text for delivery, as more specialized criteria are required in the actual production, such as appropriate wording, effective appeals, consistency between ad text and product information, and high predicted performance.

CTR is widely used to predict ad performance and make marketing strategies in advance. The CTR prediction task was first studied by Robinson et al. (2007) and has been actively researched by Rosales et al. (2012), McMahan et al. (2013), Chapelle et al. (2014), Yan et al. (2014) and Kumar et al. (2015) with traditional machine learning techniques such as logistic regression-based and factorization machines (FM)-based models (Rendle, 2010) in past decades (Yang and Zhai, 2022). In recent years, neural network (NN)-based models have performed successfully, such as the CNN-based model (Krizhevsky et al., 2012) by Niu and Hou (2020), LSTM-based model (Sak et al., 2014) by Gharibshah et al. (2020), and the combination of FM and NN model by Guo et al. (2017). While there are many methods for CTR prediction, there are few studies that evaluate the quality of ad texts at a more granular level. Coarse-level feedback may be insufficient to improve ad quality, as it is unclear where edits or what changes should be made. ADTEC, on the other hand, provides more detailed evaluation results, such as acceptability and consistency, making it easier to integrate into actual AdOps workflows.

6 Conclusion

We defined five tasks to verify the quality of ad texts and built the ADTEC, a large, versatile, and comprehensive benchmark of Japanese advertising data constructed for the NLP community, based on real-world AdOps workflows. We conducted evaluations with both PLMs and human evaluators on ADTEC to explore its characteristics, offering insights into practical workflow applications and potential areas for future improvement and research. Our findings suggest that sampling directly from real data generally benefits the model, highlighting the importance for tasks centered on natural language inference and semantic understanding. We hope that the combination of our defined tasks and datasets in this paper will advance research in ad text evaluation, bridging the fields of advertising and NLP, and paving the way for new discoveries and applications.

Limitations

Language The language is limited to Japanese as the advertising agency company is located in Japan. Expressions related to online advertising vary according to language and culture. In Japan, for example, there is a cultural preference for cluttered design (Masuda and Nisbett, 2001), which results in a lot of information being scattered around compared to the clean LPs in US and Europe. This means that different strategies and models may required for different languages and countries. Although exploring the adaptability of datasets across languages is beyond the scope of this study, it represents an exciting opportunity for future research, especially since the field of NLP in advertising remains largely unexplored.

Additionally, existing research indicates that machine translation techniques could potentially mitigate these limitations (Miyazaki and Shimizu, 2016; Masad et al., 2023; Wang and Hershcovich, 2023). Moreover, the methodology for designing these tasks and datasets is language-independent, making it applicable to other languages. Tasks such as assessing the quality of ad text (AD ACCEPTABILITY), checking the consistency of ad text with landing pages (AD CONSISTENCY), and rephrasing ad texts (AD SIMILARITY) can enhance operational efficiency across different languages. By referencing the motivations and methods outlined in this study, future researchers and engineers can design similar tasks without starting from scratch when working in different languages.

Finally, it is important to note that the primary contribution of our study is the definition of real-world advertising operations as distinct tasks and the integration of these tasks into a comprehensive benchmark, using the Japanese dataset as a case study.

Ethical Consideration

Our dataset, AdTEC, comprises documents such as ad texts, LP texts, keywords, and industry types. The dataset includes only a subset of the original data, sampled over a specific period and refined through preprocessing, as detailed in the §3.2 of our paper.

Each instance in the dataset consists of ad text data along with related information, and each instance is labeled as discussed in the task design section (§3). While the dataset is self-contained, it does not include any confidential information, as all data intended for internal use was made public through appropriate channels. However, it is important to note that the use of automatically generated ad texts by NLG models may introduce unnatural expressions, which could be considered a source of noise.

The data was collected in 2023 using both manual human curation and software programs. Authors and other full-time employees were involved in the data collection process, and they were compensated at rates above the minimum wage, and provided with basic social security benefits provided. Although no formal ethical review process was conducted, individuals involved were directly notified and consented to the data collection.

Preprocessing, cleaning, and labeling were performed on the data, and the raw data was saved for potential future uses. This dataset has already been used to evaluate ad text quality at the author's company and could be utilized for tasks such as language modeling on acceptable ad texts and hallucination detection. The dataset will be distributed to third parties via a URL and will be available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC-BY-NC-SA 4.0) International License. The authors will support and maintain the dataset, and updates may be made to correct errors or delete instances upon request. Users will be informed of any updates through our website or codebase.

Please refer to the Datasetsheets (Gebru et al., 2021), a framework designed to thoroughly cover essential information about the dataset, in Appendix A for comprehensive and detailed information.

References

Zoë Abrams and Erik Vee. 2007. Personalized ad delivery when ads fatigue: An approximation algorithm. In *Proceedings of Workshop on Internet and Network Economics*, pages 535–540.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei.

2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Olivier Chapelle, Eren Manavoglu, and Rómer Rosales. 2014. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology*, 5(4):61:1–61:34.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Chenhe Dong, Yinghui Li, Haifan Gong, Mia Xu Chen, Junxin Li, Ying Shen, and Min Yang. 2021. A survey of natural language generation. *ACM Computing Surveys*, 55:1 – 38.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Zhabiz Gharibshah, Xingquan Zhu, Arthur Hainline, and Michael Conway. 2020. Deep learning for user interest and response prediction in online display advertising. *Data Science and Engineering*, 5(1):12–26.

Konstantin Golobokov, Junyi Chai, Victor Ye Dong, Mandy Gu, Bingyu Chi, Jie Cao, Yulan Yan, and Yi Liu. 2022. DeepGen: Diverse search ad generation and real-time customization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 191–199.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: A factorization-machine based neural network for CTR prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, pages 1725–1731.

J. Weston Hughes, Keng-hao Chang, and Ruofei Zhang. 2019. Generating better search engine text advertisements with deep reinforcement learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019)*, pages 2269–2277.

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 (NAACL-HLT 2021)*, pages 255–262.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, pages 1106–1114.

Rohit Kumar, Sneha Manjunath Naik, Vani D Naik, Smita Shiralli, Sunil V.G, and Moula Husain. 2015. Predicting clicks: Ctr estimation of advertisements using logistic regression classifier. In *2015 IEEE International Advance Computing Conference (IACC)*, pages 1134–1138.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Ofri Masad, Kfir Bar, and Amir Cohen. 2023. Automatic translation of span-prediction datasets. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 160–173, Nusa Dua, Bali. Association for Computational Linguistics.

Takahiko Masuda and Richard Nisbett. 2001. Attending holistically vs. analytically: Comparing the context sensitivity of japanese and americans. journal of personality and social psychology, 81, 922-934. *Journal of personality and social psychology*, 81:922–34.

H. Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica. 2013. Ad click prediction: a view from the trenches. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013)*, pages 1222–1230.

Shaunak Mishra, Manisha Verma, Yichao Zhou, Kapil Thadani, and Wei Wang. 2020. Learning to create better ads: Generation and ranking approaches for ad creative refinement. CIKM '20, page 2653–2660. Association for Computing Machinery.

Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. 2024. Striking gold in advertising: Standardization and exploration of ad text generation.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790, Berlin, Germany. Association for Computational Linguistics.

Soichiro Murakami, Sho Hoshino, and Peinan Zhang. 2023. Natural language generation for advertising: A survey. *CoRR*, 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 (NAACL-HLT 2022)*, pages 69–78.

Tianyuan Niu and Yuexian Hou. 2020. Density matrix based convolutional neural network for click-through rate prediction. In *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 46–50.

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. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Steffen Rendle. 2010. Factorization machines. In *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pages 995–1000. IEEE Computer Society.

Helen Robinson, Anna Wysocka, and Chris Hand. 2007. Internet advertising effectiveness: the effect of design on click-through rates for banner ads. *International Journal of Advertising*, 26(4):527–541.

Rómer Rosales, Haibin Cheng, and Eren Manavoglu. 2012. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining (WSDM 2012)*, pages 293–302.

Hasim Sak, Andrew W. Senior, and Françoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, pages 338–342.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, (ICML 2017)*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, arXiv:2307.09288.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Zi Wang and Daniel Hershcovich. 2023. On evaluating multilingual compositional generalization with translated datasets. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1687, Toronto, Canada. Association for Computational Linguistics.

Alex Warstadt and Samuel R. Bowman. 2019. Grammatical analysis of pretrained sentence encoders with acceptability judgments. *arXiv preprint arXiv:1901.03438*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Ling Yan, Wu-Jun Li, Gui-Rong Xue, and Dingyi Han. 2014. Coupled group lasso for web-scale CTR prediction in display advertising. In *Proceedings of the 31th International Conference on Machine Learning (ICML 2014)*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 802–810.

Yanwu Yang and Panyu Zhai. 2022. Click-through rate prediction in online advertising: A literature review. *Information Processing and Management: an International Journal*, 59(2).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Datasheets

We adhere to the existing dataset documentation framework proposed by Gebru et al. (2021) to provide comprehensive information about our dataset.

A.1 Motivation

For what purpose was the dataset created? This dataset is designed to evaluate the quality of ad texts in multiple aspects from the perspective of practical advertising operations.

Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)? The authors of this paper created the dataset and conducted the research.

Who funded the creation of the dataset? N/A

A.2 Composition

What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? The dataset represents documents, e.g., ad texts, LP texts, keywords, and industry types.

How many instances are there in total (of each type, if appropriate)? See Table 5 and 4.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? As described in 3.2, the dataset includes only a subset of the original data, sampled over a specific period and refined through preprocessing.

What data does each instance consist of? Our dataset primarily comprises ad text data and related information, including LP texts, keywords, and industry types.

Is there a label or target associated with each instance? Yes. See Section 3.

Is any information missing from individual instances? N/A.

Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? N/A.

Are there recommended data splits (for example, training, development/validation, testing)? Yes. See Section 3.2 and Table 5.

Are there any errors, sources of noise, or redundancies in the dataset? The use of automatically generated ad texts by NLG models may introduce unnatural expressions, which can be considered a form of noise.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? The dataset is self-contained for the tasks described in the paper.

Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? Data originally intended for internal use but subsequently made public through appropriate channels.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

Does the dataset identify any subpopulations (for example, by age, gender)? No.

Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? No.

Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? No.

A.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/ derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? See Section 3.2.

What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation,

software programs, software APIs)? See Section 3.2.

If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)? As described in Section 3.2, we utilized data from a specific period without employing probabilistic sampling. Items that could not be technically included in the dataset were removed.

Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)? Authors and other full-time employees. All workers involved in this research are employed as full-time employees by the company and are compensated at rates above the minimum wage set by local authorities. Additionally, they are provided with basic social security benefits.

Over what timeframe was the data collected? The data collection was conducted in 2023.

Were any ethical review processes conducted (for example, by an institutional review board)? No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)? In question directly.

Were the individuals in question notified about the data collection? Yes.

Did the individuals in question consent to the collection and use of their data? Yes.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? N/A.

A.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? Yes. See Section 3.2.

Was the "raw" data saved in addition to the preprocessed/cleaned/ labeled data (for example, to support unanticipated future uses)? Yes.

Is the software that was used to preprocess/clean/label the data available? No.

A.5 Uses

Has the dataset been used for any tasks already? Yes. The dataset has been used to evaluate the quality of ad texts at the author's company.

Is there a repository that links to any or all papers or systems that use the dataset? Yes. <URL>

What (other) tasks could the dataset be used for? Various tasks using the dataset can be considered, such as language modeling on the acceptable ad texts and hallucination detection with the AD CONSISTENCY dataset.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? As mentioned in Section 4, the AD PERFORMANCE ESTIMATION task masks certain entities to protect information. Consequently, users may need to complete the task with incomplete information, which could affect its usability.

Are there tasks for which the dataset should not be used? N/A.

A.6 Distribution

Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? Yes.

How will the dataset be distributed (for example, tarball on website, API, GitHub)? <URL>.

When will the dataset be distributed? After this paper is accepted, as soon as possible.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The dataset is distributed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License.

Name	URL
Tohoku BERT	tohoku-nlp/bert-{base,large}-japanese-v2
Waseda RoBERTa	nlp-waseda/roberta-{base,large}-japanese-with-auto-jumanpp
XLM-RoBERTa	xlm-roberta-{base,large}
CALM2	cyberagent/calm2-7b-chat
ELYZA	elyza/ELYZA-japanese-CodeLlama-7b-instruct
OpenAI	2023-08-01-preview

Table 8: URLs and names of used PLMs.

Model	PreTokenizer (Dictionary)	Tokenization Unit	Dataset	#Vocab	#Param
Tohoku BERT$_{\text{BASE}}$	MeCab (IPADic+NEologd)	BPE	Wikipedia (Ja)	32K	111M
Tohoku BERT$_{\text{LARGE}}$	MeCab (IPADic+NEologd)	BPE	Wikipedia (Ja)	32K	337M
Waseda RoBERTa$_{\text{BASE}}$	Juman++	Unigram LM	Wikipedia (Ja) + CC (Ja)	32K	110M
Waseda RoBERTa$_{\text{LARGE}}$	Juman++	Unigram LM	Wikipedia (Ja) + CC (Ja)	32K	336M
XLM-RoBERTa$_{\text{BASE}}$	—	Unigram LM	Multilingual CC	250K	278M
XLM-RoBERTa$_{\text{LARGE}}$	—	Unigram LM	Multilingual CC	250K	559M

Table 9: List of trained language models used in the experiment, where CC and Ja represents CommonCrawl and Japanese data, respectively.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? The third party data is not intended for commercial use and is subject to the organization's terms and conditions.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No

A.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The authors will be.

How can the owner/curator/ manager of the dataset be contacted (for example, email address)? By the email address.

Is there an erratum? N/A.

Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? It is possible to update the dataset on the website or codebase to correct errors or delete instances upon request.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? N/A.

Will older versions of the dataset continue to be supported/hosted/ maintained? It depends on the nature of the dataset update. We may inform dataset users through our website or codebase.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? No. We consider that the dataset should be used exclusively for our task. Any further work should be a separate contribution from ours.

B Additional Information in Dataset Construction

B.1 Annotation Guidelines

The annotation guidelines for AD ACCEPTABILITY, AD CONSISTENCY, and AD SIMILARITY are shown in Figure 7 and 8.

B.2 Label Distributions

The label distributions for AD ACCEPTABILITY, AD CONSISTENCY, A^3 RECOGNITION, AD SIMILARITY, and AD PERFORMANCE ESTIMATION tasks are shown in Table 11 and Figure 9.

C Additional Information in Experiment

C.1 Model Details

The versions and URLs of the models we used are listed in Table 8. The detailed settings for the pretraining of fine-tuned models, such as the tokenizer and dataset, are provided in Table 9.

Parameter	Value
Fine-tuned Encoder Models	
Lerning Rate	{2e-5, 5.5e-5, 2e-6}
Seed	0
Epoch	30
Early Stopping Patience	3
Optimizer	Adam
Max Sequence Length	128
Zero-/Few-shot LLMs	
Attempts Per an Instance	5
Temperature	0.8
Max New Tokens	64
Fine-tuned LLMs	
LoRA alpha	16
LoRA dropout	0.1
Bottleneck r	16

Table 10: Hyperparameters used to the fine-tuned encoder and LLM evaluators. Numbers in curly brackets represent the range of possible values.

Following the instruction: `{task_summary}`

 # Judgment Criteria (In the A³ RECOGNITION task, a list of descriptions of the A³ is provided.)
`{criteria}`

 # Format instruction
`{format}`

`{input}`
output:

Figure 6: The prompt template for the LLM experiment. The text in italics and curly brackets is the prompt text and the placeholders, respectively. The number of inputs is equal to n in the n-shot settings.

C.2 Hyperparameters

We used PyTorch (Paszke et al., 2019) and HuggingFace (Wolf et al., 2020) libraries for the model implementation. We present the hyperparameters employed during the training of both fine-tuned encoder models and zero-/few-shot settings on LLMs. Table 10 details the specific hyperparameters used in our experiments.

C.3 Prompts

The template for the prompt in each task is illustrated in Figure 6. We begin by outlining the purpose and overview of the task, followed by an explanation of the criteria and standards for evaluation. Next, we provide instructions on the response format, and in some cases, additional few-shot examples are included. Tables 13 and 14 show the prompts actually used in the tasks. Terms in curly brackets are variables to be filled in. All variables, except `{few_shot_examples}`, correspond to the fields described in Section 3. Note that all prompts are translated from Japanese into English for the presentation brevity.

C.4 Evaluators

Fine-tuning Setting with Encoder Models. In the fine-tuning settings, models receive ad text and other values $\boldsymbol{x} = (x_i)_{i=1}^{|\boldsymbol{x}|}$ as input and predict labels $\boldsymbol{y}$, where x_i represents the i-th token of the input sequence. The hyperparameters used are shown in Appendix Table 10. AD ACCEPTABILITY and AD CONSISTENCY involve predicting binary label $\boldsymbol{y} \in \{0, 1\}$ using an MLP from the encoded vector representation $h^{[\texttt{CLS}]}$, where [CLS] is the special token for classification. The inputs of the AD ACCEPTABILITY task only use the ad text $\boldsymbol{x}^{\text{ad}}$, while AD CONSISTENCY uses $\boldsymbol{x}^{\text{ad}} \oplus \boldsymbol{x}^{\text{LP}}$ as input, where $\boldsymbol{x}^{\text{LP}}$ and $\oplus$ represent the LP text and concatenation by a special token [SEP], respectively. A^3 RECOGNITION involves predicting all possible labels from a given ad text, $\boldsymbol{x}^{\text{ad}} = (x_i^{\text{ad}})_{i=1}^{|\boldsymbol{x}^{\text{ad}}|}$. We adopted the doc-based architecture described in Murakami et al. (2022), where the encoder model was used to obtain the vector representation $h^{[\texttt{CLS}]}$. This vector was then fed into an MLP layer, producing a label probability distribution $\boldsymbol{m} = \texttt{Sigmoid}(\texttt{MLP}(h^{[\texttt{CLS}]}))$, where $\boldsymbol{m} = (m_k)_{k=1}^{K}$ and K is 21. This number corresponds to the number of A^3 labels defined by Murakami et al. (2022), as shown in Appendix Table 11. AD SIMILARITY and AD PERFORMANCE ESTIMATION are regression tasks that involve predicting a value of range $\boldsymbol{y} \in [1, 5]$ and $\boldsymbol{y} \in [0, 100]$, respectively. Similar to the text classification task, we concatenated all inputs for each task with the special token [SEP], encoded the input into vector representation $h^{[\texttt{CLS}]}$, and then fed the vector into the MLP layer to predict $\boldsymbol{y}$.

Zero-/Few-shot Setting with LLMs. Before conducting the experiment, we sampled 100 instances from the development set to adjust the parameters and prompts. The prompts we used are shown in Figure 6. Additionally, we determined the number of shots that performed best in the

Evaluator	AD ACCEPT. Accuracy/F1-score	AD CONSIST. Accuracy/F1-score	AD PERFORM. EST. Pearson/Spearman	A^3 RECOGNITION F1-micro/-macro	AD SIMILARITY Pearson/Spearman
			Fine-tuned Encoder Models		
Tohoku $BERT_{BASE}$	0.685/0.691	0.757/0.504	0.437/0.454	0.753/0.629	0.769/0.803
Tohoku $BERT_{LARGE}$	0.711/0.688	**0.767**/0.552	**0.480/0.497**	0.774/**0.694**	0.773/0.807
Waseda $BERT_{BASE}$	0.615/0.639	0.725/0.388	0.444/0.454	0.641/0.442	0.749/0.797
Waseda $BERT_{LARGE}$	0.598/0.637	0.755/0.474	0.445/0.457	0.663/0.517	0.740/0.800
XLM-$RoBERTa_{BASE}$	0.694/0.677	0.743/0.465	0.425/0.439	0.730/0.542	0.846/0.870
XLM-$RoBERTa_{LARGE}$	0.705/0.690	0.758/0.519	0.453/0.457	**0.778**/0.677	**0.878/0.878**
			Zero-/Few-shot LLMs		
$CALM2_{7b}$	0.520/0.115	0.381/0.472	0.006/0.013	0.154/0.042	0.036/0.036
$ELYZA_{7b}$	0.352/0.520	0.628/0.771	0.003/0.046	0.196/0.044	0.015/-0.004
GPT-3.5	0.369/0.489	0.528/0.570	-0.013/-0.022	0.255/0.064	0.389/0.385
GPT-4	0.325/0.433	0.583/0.612	0.028/0.073	0.417/0.113	0.776/0.811
			Fine-tuned LLMs		
$ELYZA_{7b}$	0.638/0.638	0.692/0.694	0.240/0.235	0.379/0.280	0.684/0.740
Human	**0.732/0.790**	0.703/**0.807**	—	0.564/0.538	0.699/0.765

Table 12: Complete version of Table 6 which showing the performance of PLMs and human evaluators on the test set. Bold indicates the best result across all methods.

development set: 3-shot for AD SIMILARITY and AD PERFORMANCE ESTIMATION task, 2-shot for A^3 RECOGNITION, and zero-shot for the others. We calculated the final score by averaging five runs with different few-shot examples for a single instance.

Fine-tuning Setting with LLM. We fine-tuned ELYZA-7b, one of the open-sourced LLMs used in our study. We utilize QLoRA tuning, which applies the query and value of the attention module, training it for one epoch with the hyperparameters provided in Table 10.

C.5 Additional Results and Discussion on Fine-Tuning LLM Settings

The results of this fine-tuning are detailed in the Table 12. By fine-tuning $ELYZA_{7b}$, we observed performance improvements in most tasks as expected. Although it still falls short of fine-tuned encoder models, it outperformed GPT-4 in AD ACCEPTABILITY and AD CONSISTENCY tasks. While it did not match GPT-4's performance in A^3 RECOGNITION and AD SIMILARITY tasks, it achieved performance levels close to human capabilities in AD SIMILARITY.

The extent of improvement varied across tasks. Notably, in AD CONSISTENCY, the accuracy increased by only 0.064, whereas in AD SIMILARITY, the Pearson and Spearman correlations improved dramatically to 0.659 and 0.736, respectively. The AD ACCEPTABILITY and A^3 RECOGNITION tasks saw moderate improvements. This suggests that pre-training is particularly effective for tasks similar to AD CONSISTENCY, likely because the pre-training data, derived from web crawls, includes general knowledge applicable to AD CONSISTENCY. Conversely, the low scores in zero-/few-shot experiments for AD SIMILARITY indicate that applying general knowledge from pre-training is challenging. However, fine-tuning appears to effectively map the pre-trained knowledge to the specific task, significantly enhancing performance. For AD ACCEPTABILITY and A^3 RECOGNITION tasks, although additional training led to some improvement, the performance remained relatively low, suggesting that the provided training data was insufficient to fully grasp the knowledge required for these tasks.

Label	Train	Dev	Test	Total
AD ACCEPTABILITY				
Acceptable	15,099	–	850	15,949
Not acceptable	20,278	–	1,150	21,428
Total	35,377	–	2,000	37,377
AD CONSISTENCY				
Consistent	8,708	–	620	9,328
Not consistent	20,048	–	1,380	21,428
Total	28,756	–	2,000	30,756
A^3 RECOGNITION				
Special deals	230	63	42	335
Discount price	80	22	16	118
Reward points	50	17	14	81
Free	283	66	62	411
Special gift	83	20	18	121
Features	746	178	190	1,114
Quality	35	17	11	63
Problem solving	10	4	3	17
Speed	95	23	20	138
User-friendliness	213	65	32	310
Transportation	53	17	10	80
Limited offers	36	10	6	52
Limited time	43	12	5	60
Limited target	81	10	18	109
First-time limited	16	5	4	25
Performance	47	13	9	69
Largest/No.1	108	11	20	139
Product lineup	167	38	42	247
Trend	69	15	13	97
Others	117	33	23	173
Story	73	16	9	98
Total	2,635	655	567	3,857
A^3 RECOGNITION (#Labels per document)				
0	337	94	84	515
1	769	198	165	1,132
2	485	98	100	683
3	182	44	47	273
4	69	26	10	105
5	10	5	3	18
6	4	0	1	5
Total	1,856	465	410	2,731
AD SIMILARITY				
$1 \leq x < 2$	527	66	67	660
$2 \leq x < 3$	845	105	108	1,058
$3 \leq x < 4$	2,739	343	344	3,426
$4 \leq x < 5$	790	99	100	989
$5 \leq x$	79	10	10	99
Total	4,980	623	629	6,232

Table 11: Label distribution of AD ACCEPTABILITY, AD CONSISTENCY, A^3 RECOGNITION and AD SIMILARITYtask. Note that the number of labels in A^3 RECOGNITION does not necessarily equal the number of sentences, because a single document can have multiple labels.

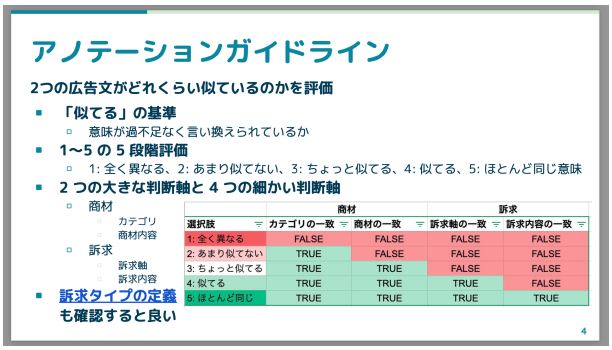

Figure 7: Annotation guideline for AD ACCEPTABILITYand AD CONSISTENCY task.

Figure 8: Annotation guideline for AD SIMILARITY task.

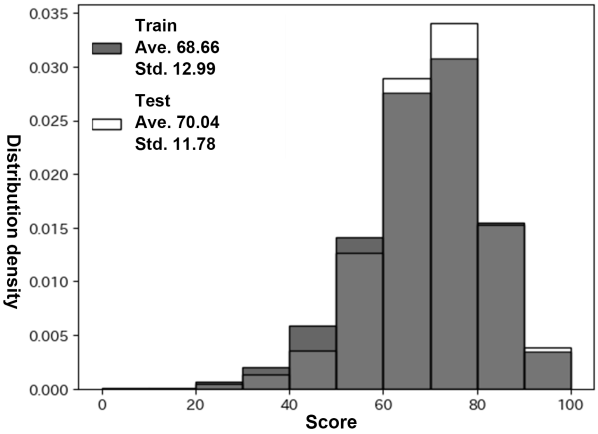

Figure 9: Label distribution of AD PERFORMANCE ESTIMATION.

AD ACCEPTABILITY
Based on the following settings, determine whether the input sentence is acceptable as an advertisement expression. # Criteria - Output *acceptable* if the sentence is coherent, fluent, and easy to read. - Output *unacceptable* if the sentence is unnatural due to unnecessary and excessive repetition, inappropriate use of symbols, etc. # Formats - Answer with either *acceptable* or *unacceptable*. - Do not output anything other than your answer. `{few_shot_examples}` Input sentence: `{ad_text}` Output:
AD CONSISTENCY
Based on the following settings, determine whether the given input sentence is consistent with the LP text. # Criteria - Output *inconsistent* if the input sentence contains expressions not included in the LP text. - Output *inconsistent* if the input sentence outputs different numbers or names from those mentioned in the LP text. # Formats - Answer with either *consistent* or *inconsistent*. - Do not output anything other than your answer. `{few_shot_examples}` Input sentence: `{ad_text}` LP text: `{lp_text}` Output:
AD PERFORMANCE ESTIMATION
Based on the following settings, estimate the performance of the given advertisement information. # Criteria - The good quality of the ad text (whether it is attractive, whether it is effective in the industry) - The relevance of keywords and ad text - The relevance of LP text and ad text # Formats - Do not output anything other than your answer. - Answer with a number between 0 and 100. `{few_shot_examples}` Industry: `{industry_type}` Keyword: `{keyword}` Title: `{title_1}` `{title_2}` `{title_3}` Description: `{description_1}` `{description_2}` Output:
AD SIMILARITY
Based on the following settings, rate the advertising similarity of the given two input sentences. # Criteria - Judge based on the similarity in product category/product name and the similarity in appeal axis/persuasive expressions. - Closer to 5 if similar, closer to 1 if not similar. # Formats - Do not output anything other than your answer. - Answer with a decimal number between 1 and 5. `{few_shot_examples}` Input sentence 1: `{ad_text_1}` Input sentence 2: `{ad_text_2}` Output:

Table 13: Prompts used in AD ACCEPTABILITY, AD CONSISTENCY, AD PERFORMANCE ESTIMATION, and AD SIMILARITY tasks.

A^3 RECOGNITION

Based on the following settings, list all the aspects of advertising appeals included in the given input sentence. Choose aspect labels from the following list. The list is in the format of "Label": "Description".

List of aspects of advertising appeals
- *Special deals*: Expressions emphasizing a sense of value such as price or discounts
- *Discount price*: Expressions emphasizing specific prices or discounts
- *Reward points*: Expressions emphasizing the rebate of points or money
- *Free*: Expressions emphasizing that something is free
- *Special gift*: Expressions emphasizing that perks are included
- *Features*: Expressions emphasizing the content or features of the service
- *Quality*: Expressions emphasizing high quality or a high-grade feel
- *Problem solving*: Expressions emphasizing the solution to customers' problems
- *Speed*: Expressions emphasizing the speed of delivery or procedures
- *User-friendliness*: Expressions emphasizing the ease of use of the service
- *Transportation*: Expressions emphasizing good accessibility
- *Limited offers*: Expressions emphasizing some form of limitation
- *Limited time*: Expressions emphasizing a limited period for service provision
- *Limited target*: Expressions emphasizing that the service is provided to/for a limited target
- *First-time limited*: Expressions emphasizing that the service is provided only for the first time
- *Performance*: Expressions emphasizing the achievements of the service or company
- *Largest/No.1*: Expressions emphasizing the scale or being No. 1 of the service or company
- *Product lineup*: Expressions emphasizing the assortment or the number of stores
- *Trend*: Expressions emphasizing that it is trending or in vogue
- *Other*: Persuasive expressions suitable for advertising that do not fall into the above labels
- *Story*: Expressions emphasizing the synopsis of the work
- *No Match*: Label for when there is no persuasive expression

Formats
- Output each aspects separated by "|"
- Do not output anything other than your answer
- If no aspects apply, output "*No Match*"

`{few_shot_examples}`
Input sentence: `{ad_text}`
Output:

Table 14: Prompt used in A^3 RECOGNITION task.