AdParaphrase v2.0: Generating Attractive Ad Texts Using a Preference-Annotated Paraphrase Dataset

Soichiro Murakami[1], **Peinan Zhang**[1], **Hidetaka Kamigaito**[2,3], **Hiroya Takamura**[3], **Manabu Okumura**[3]
[1]CyberAgent, Inc., [2]Nara Institute of Science and Technology, [3]Institute of Science Tokyo
{murakami_soichiro,zhang_peinan}@cyberagent.co.jp,
kamigaito.h@is.naist.jp, {takamura,oku}@pi.titech.ac.jp

Abstract

Identifying factors that make ad text attractive is essential for advertising success. This study proposes AdParaphrase v2.0, a dataset for ad text paraphrasing, containing human preference data, to enable the analysis of the linguistic factors and to support the development of methods for generating attractive ad texts. Compared with v1.0, this dataset is 20 times larger, comprising 16,460 ad text paraphrase pairs, each annotated with preference data from ten evaluators, thereby enabling a more comprehensive and reliable analysis. Through the experiments, we identified multiple linguistic features of engaging ad texts that were not observed in v1.0 and explored various methods for generating attractive ad texts. Furthermore, our analysis demonstrated the relationships between human preference and ad performance, and highlighted the potential of reference-free metrics based on large language models for evaluating ad text attractiveness. The dataset is publicly available at: https://github.com/CyberAgentAILab/AdParaphrase-v2.0.[1]

1 Introduction

Advertisements play a vital role in marketing, raising awareness of products or services, capturing user interests, and driving actions such as clicks. To maximize their effectiveness, ad writers must create attractive ad texts that appeal to users. However, with the growing demand for online advertising, manual ad text creation is reaching practical limitations, highlighting the need for automatic ad text generation (ATG) (Murakami et al., 2023). Writing attractive ad texts requires considering two aspects: *what-to-say* (the content to be advertised, such as price or product name) and *how-to-say* (the way the content is expressed). This study focuses on the *how-to-say* aspect in exploring methods for generating attractive ad texts, aiming to identify linguistic factors that capture the user's interest.

	Ad Text	#Pref
(a)	*Up to 50% discount on your first purchase*	0
	Get up to 50% off on your first purchase	9
(b)	*The industry's lowest prices*	3
	Top-class low prices in the industry	7

Table 1: Example of AdParaphrase v2.0, translated into English for visibility. #Pref represents the number of evaluators who preferred each ad text. Those who chose "skip" are not included.

Many studies have investigated the factors that influence ad performance and human preference (Youngmann et al., 2020; Yuan et al., 2023). However, identifying the linguistic factors presents a significant challenge because of the intricate interplay between the semantic content and its linguistic expression. A clear analysis of the linguistic factors requires disentangling them and focusing exclusively on their impact (Pryzant et al., 2018).

To address this challenge, Murakami et al. (2025) introduced AdParaphrase, which is a dataset comprising paraphrase pairs of ad texts, annotated with human preferences from ten evaluators. By controlling the content, the dataset allows us to investigate how linguistic expressions alone affect the attractiveness of the ad. Using this dataset, they identified linguistic factors, such as noun count, that significantly affect human preferences. In addition, they demonstrated that these findings can improve the generation of attractive ad texts.

However, the small size of their dataset, AdParaphrase, presents notable limitations. The dataset contains only 725 paraphrase pairs created by professional ad writers and is insufficient for conducting comprehensive and reliable analyses or training ATG models. Consequently, previous studies have primarily relied on in-context learning (ICL) (Brown et al., 2020), leaving other promising

[1]The dataset is provided under the CC BY-NC-SA 4.0 license.

Findings of the Association for Computational Linguistics: ACL 2025, pages 15212–15230

approaches, such as preference tuning (Rafailov et al., 2023), unexplored.

To address these limitations, we present AdParaphrase v2.0, an expanded version of the original dataset, with referring to the original dataset as v1.0. Table 1 presents paraphrase examples from the dataset. The number of paraphrase pairs annotated with human preferences in v2.0 is approximately 20 times larger than v1.0. This expansion enables a comprehensive analysis and encourages the exploration of other ATG approaches. The dataset was built using scalable methods including large language models (LLMs) and crowdsourcing, with manual annotations for paraphrase identification and preference judgment.

In the experiments, we analyzed AdParaphrase v2.0 and identified multiple linguistic factors influencing human preferences that were not identified in v1.0 (§5.1). We then evaluated various methods for generating attractive ad texts, including ICL, instruction tuning, and preference tuning, by examining the characteristics of each approach (§5.2). In addition, our analysis identified the relationships between human preferences and ad performances, and demonstrated the suitability of reference-free metrics for the automatic evaluation of ad text attractiveness (§6). We hope AdParaphrase v2.0 will drive further advancements in generating attractive ad texts.

2 Related Work

2.1 Ad Text Optimization

Optimizing ad texts to enhance ad performance is a critical challenge for advertisers. To this end, various approaches have been developed such as ATG and text analysis (Murakami et al., 2023).

ATG approaches are broadly classified into two categories: generation from scratch (Bartz et al., 2008; Hughes et al., 2019) and ad text refinement (Youngmann et al., 2020; Murakami et al., 2025). The former involves creating ad text from sources, such as keywords and landing pages (Kamigaito et al., 2021; Mita et al., 2024), whereas the latter focuses on improving existing ad texts (Mishra et al., 2020). This study falls into the latter category.

Using text analysis, previous studies investigated factors affecting attractiveness, such as persuasion strategies (Yuan et al., 2023), emotions (Youngmann et al., 2020), and advertising appeal (Murakami et al., 2022). The key difference between previous studies and our work is that we focus on the attractiveness of linguistic expression in ad texts. The factors that influence attractiveness can be broadly divided into *what-to-say* and *how-to-say*. Although previous studies often focused on *what-to-say* without explicitly distinguishing between the two, we specifically focus on *how-to-say*.

2.2 Paraphrase Generation

Our study is closely related to paraphrase generation, as it focuses on rephrasing ad texts into more attractive expressions while preserving their meaning. Paraphrase generation has long been a central challenge in natural language processing, with numerous datasets and methods proposed across various domains (Zhou and Bhat, 2021).

This study differs from previous studies in two key aspects: First, it targets ad texts, a domain with unique characteristics distinct from previously studied areas such as social media (Lan et al., 2017) and questions (Zhang et al., 2019). Second, it prioritizes human preference in paraphrase pairs, specifically examining linguistic expressions that enhance the attractiveness of ad texts—a perspective unique to the advertising domain. We hope that our dataset will expand the scope of future research on paraphrase generation.

3 Method of Dataset Construction

This section describes the design principles of AdParaphrase v2.0 (§3.1), the three-step construction process involving paraphrase candidate collection (§3.2), paraphrase identification (§3.3), and preference judgment (§3.4), and the quality control measures implemented throughout its construction (§3.5).

3.1 Principles of Dataset Design

Our design principles are threefold: (1) ensuring that the dataset is large enough to support both analysis and model training; (2) incorporating a diverse range of paraphrasing cases; and (3) making the dataset publicly available under a proper license for research purposes.

To achieve Principle (1), over 10,000 data samples were collected. This quantity was determined based on the benchmarks and requirements observed in previous studies (Jha et al., 2023; Mita et al., 2024) for reliable data analysis and effective model training. To address Principle (2), a wide range of paraphrased expressions were covered beyond simple phenomena such as "word

order changes" by providing explicit stylistic instructions during paraphrase generation. Finally, in line with Principle (3), the dataset was constructed using methods compatible with open distribution for research purposes. Specifically, we leveraged crowdsourcing and utilized open LLMs whose licenses permit the redistribution of the generated content.

3.2 Collecting Paraphrase Candidates

ADPARAPHRASE V2.0 was constructed based on CAMERA (Mita et al., 2024), a publicly available Japanese ad text dataset for ATG. In this study, by leveraging all ad texts from the dataset as source texts, we collected paraphrase pairs by generating their paraphrases using both LLMs and crowdworkers. While the quality would be ensured by relying solely on professional ad writers to create paraphrases, it is impractical to construct large-scale datasets with the method because of resource constraints. To address this issue, we leveraged 133 high-quality paraphrase pairs from ADPARAPHRASE V1.0 created by professional ad writers as references for LLMs and crowdworkers. This approach combines the expertise of professional writers with automated methods to efficiently generate numerous paraphrase candidates. The procedure for generating paraphrases using LLMs and crowdworkers is as follows:

Paraphrase Generation using LLMs Paraphrase candidates were generated using LLMs, known for their paraphrase-generation capabilities (Cegin et al., 2023), via In-Context Learning (ICL) (Brown et al., 2020). For this approach, high-quality paraphrase examples from professional writers were provided as few-shot examples, along with instruction texts as prompts. To enhance the diversity of paraphrases in accordance with Principle (2), stylistic instructions were also incorporated into the prompts. We defined 40 types of stylistic instructions, such as *"Use simpler syntax"*, to guide LLMs in generating paraphrase candidates based on specified styles.[2] Stylistic instructions were randomly selected for each ad text. Examples of prompts and stylistic instructions are provided in Appendix A. Moreover, multiple LLMs with different training datasets and model sizes were used. The selection of LLMs was based on Principle (3) and whether they were pre-trained on Japanese corpora. Specifically, we selected four models.[3] For example, `Swallow-70B` is a model based on Llama 3.1 and is distributed under the Llama 3.1 license,[4] which permits the use of model-generated texts for research purposes, including model training.

Paraphrase Generation by Crowdworkers We used a crowdsourcing service.[5] The same instructions and paraphrase examples as those given to the LLMs were provided to the crowdworkers as annotation guidelines. Because most workers lack experience in creating ad texts, additional knowledge about ad text creation (e.g., *"Include words that encourage action"*) was also included in the guidelines. The complete guidelines are available in Appendix A.

3.3 Paraphrase Identification

Manual labeling was conducted to indicate whether the generated candidates are really a paraphrase at the sentence level. To reduce the manual labor, we first applied rule-based filtering to exclude (1) candidates that are clearly not a paraphrase (e.g., contain different dates or monetary amounts) and (2) ad texts exceeding 30 characters. The length constraint was based on guidelines from ad platforms such as Google Ads[6] because texts beyond this limit cannot be delivered. Paraphrase identification (PI) was conducted via crowdsourcing, whereby five workers evaluated each ad text pair and made a binary judgment on whether it qualifies as a paraphrase. The final label for each pair was determined by majority vote. The instructions provided to the workers and example paraphrase pairs are presented in Appendix B and D, respectively.

3.4 Human Preference Judgment

Preference judgments were conducted for valid paraphrase pairs via crowdsourcing, with each pair judged by ten workers. Workers were asked to select the more attractive ad text or "skip" if both were equally attractive. To address the subjective nature of preference judgments, we followed the guidelines of Wang et al. (2021) and provided

[2]The results from our analysis of the effect of stylistic instructions are provided in Appendix E. We confirmed that explicitly specifying stylistic instructions enables the generation of lexically and syntactically diverse paraphrase candidates.

[3]The four models include tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1, tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.1, cyberagent/calm2-7b-chat, and cyberagent/calm3-22b-chat on Hugging Face Hub (Wolf et al., 2020).

[4]https://www.llama.com/llama3_1/license/

[5]https://crowdsourcing.yahoo.co.jp/

[6]https://ads.google.com

Model	#Generated	#Filtered	#Para.	Pass Rates PI	Pass Rates Pref
CALM2-7B	16,365	2,107	1,173	7.2	22.9
CALM3-22B	16,365	6,287	4,551	27.8	21.4
Swallow-8B	16,365	4,942	3,623	22.1	20.9
Swallow-70B	16,365	5,226	4,174	25.5	19.5
Crowd worker	5,000	3,775	2,939	**58.8**	**25.8**
Total	70,460	22,337	16,460	23.4	21.7

Table 2: Statistics of AdParaphrase v2.0. #Generated, #Filtered, #Para. refer to the number of generated paraphrase candidates, the number of paraphrase candidates that passed the rule-based filtering, and the number of valid paraphrases judged by a majority of workers, respectively. PI and Pref stand for the pass rates of paraphrase identification and preference judgment.

the workers with multiple aspects of attractiveness, such as *"more clickable?"* and *"easier to understand"* as well. The complete annotation guidelines are provided in Appendix C.

3.5 Quality Control

Several measures were implemented to ensure high annotation quality despite inherent biases, such as positional bias (Wang et al., 2024). Positional bias was mitigated by randomizing the order of the options presented to the workers. In addition, attention checks (Klie et al., 2024) were included using identical ad text pairs with predefined correct answers (e.g., *paraphrase* for the PI task and *skip* for preference judgment), rejecting responses from annotators failing these checks to maintain quality.

4 Dataset Statistics and Analysis

4.1 Dataset Statistics

Table 2 summarizes the dataset statistics obtained for the paraphrase construction process described in §3. First, for paraphrase candidate collection, 16,365 ad texts from CAMERA were used as inputs, obtaining 70,460 paraphrase candidates through four LLMs and crowdsourcing. As source text, crowdworkers used 5,000 texts randomly sampled from CAMERA. Second, rule-based filtering was applied, resulting in 22,337 paraphrase candidates. Many candidates were removed during this filtering step primarily because they exceeded the length constraints. Third, 16,460 candidates were judged as paraphrases in PI. Finally, conducting preference judgments on the identified paraphrase pairs yielded 16,460 pairs of preference judgment data.

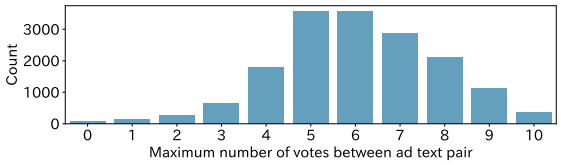

Figure 1: Distribution of maximum number of votes between ad text pair in preference judgments.

4.2 Inter-Annotator Agreement

Inter-annotator agreement (IAA) for PI (§3.3) and preference judgment (§3.4) was measured using Fleiss' kappa (Fleiss et al., 1971). The kappa value for PI was 0.442, indicating moderate agreement, whereas that for preference judgment was 0.167, indicating slight agreement (Landis and Koch, 1977). The relatively low agreement in preference judgment likely reflects the subjective nature, which is consistent with the results of previous studies on ad text evaluation (Mita et al., 2024).

4.3 Evaluation of Paraphrase Candidates

Table 2 presents the pass rates for PI and preference judgment across different models. The pass rate for PI represents the proportion of generated texts that passed both rule-based filtering and manual annotation, whereas the pass rate for preference judgment indicates the proportion of paraphrases judged as attractive by at least eight evaluators.

For PI, crowdworkers achieved the highest pass rate, and larger LLMs such as CALM3-22B performed better. In preference judgment, crowdworkers again outperformed LLMs, with 25.8% of their paraphrases judged as attractive. Among LLMs, CALM2-7B showed a slightly higher rate. The gap between LLMs and crowdworkers in preference judgment was small, suggesting that LLMs, despite slightly underperforming humans, are still effective for generating attractive paraphrases.

4.4 Distribution of Preference Judgments

The histogram showing the distribution of preference judgment results is presented in Figure 1. The x-axis represents the number of evaluators who preferred the same ad text, excluding "skip" responses. For example, a value of seven indicates that seven out of ten evaluators preferred the same ad text, whereas zero indicates that all evaluators skipped it.

The distribution of preference judgments and their IAA (§4.2) revealed an inconsistency in human preference for ad text paraphrase pairs. Specif-

Ver.	Labels	#Pairs	Pref.	Training	Creators
v2.0	Para	16,460	✓	Allowed	Crowdworker, Open LLMs
	Non-Para	5,877	–		
v1.0	Para	725	✓	Limited	Ad writers, Closed LLMs
	Non-Para	513	–		

Table 3: Comparison of ADPARAPHRASE v1.0 and v2.0.

ically, the most common agreement level involved five to six evaluators. However, 3,570 cases, with at least eight evaluators preferring the same ad text, showed moderate agreement with an IAA of 0.480, measured by Fleiss' kappa. This non-random agreement level, which was particularly noticeable in cases of strong preference, suggests that differences in linguistic expressions are likely to influence human preferences.

4.5 Dataset Comparison

Table 3 compares ADPARAPHRASE v1.0 and v2.0. ADPARAPHRASE v2.0 includes over 20 times more paraphrases compared to v1.0. Furthermore, our dataset adheres to Principle (3), in that it is freely available for research, including model training. In contrast, v1.0 relies on GPT-3.5 and GPT-4 via the Azure OpenAI API, that imposes licensing restrictions that limit its usability.[7]

5 Experiments

Through dataset construction, we collected ad text pairs with human preference annotations that were 20 times larger in scale than those in v1.0. Using the dataset, we conducted two experiments: (1) an analysis of linguistic features influencing human preferences and (2) an ATG task. The first experiment leveraged our larger dataset to identify the linguistic features influencing human preferences that were not revealed in v1.0. The second experiment evaluated the effectiveness of recent text-generation techniques, such as instruction tuning (Wei et al., 2022) and preference tuning (Rafailov et al., 2023), for the ATG task. This extends the prior work limited to ICL. Through the experiment, we assessed the potential of these methods for generating more attractive ad texts.

[7] https://azure.microsoft.com/en-us/products/ai-services/openai-service/

5.1 Analysis of Linguistic Features

In this experiment, we focused on 3,570 paraphrase pairs with moderate preference agreement (§4.4), analyzing how differences in linguistic expressions influence preferences using a chi-square test.

5.1.1 Experimental Settings

Linguistic Features The objective of ad texts is to capture people's attention and draw their interest. Thus, factors such as visibility, informativeness, and readability play a crucial role in enhancing their attractiveness (Wang and Pomplun, 2012; Schwab, 2013). We analyzed how linguistic features related to expression and style influence human preferences. Following Murakami et al. (2025), linguistic features were categorized into four groups: *raw text*, *lexical*, *syntactic*, and *stylistic*. A list of the linguistic features is presented in Table 4.[8] As a raw text feature, we used character count, which affects the informativeness and readability of the text. The lexical features include the number of content words, character types, and lexical choice. Content words are related to informativeness, whereas character types are associated with readability (Sato et al., 2008). Lexical choice was measured by counting common and proper nouns, assuming that commonly used words are preferred. Syntactic features measure text complexity and fluency, including the depth of the dependency tree, the dependency link length, and perplexity (PPL). Stylistic features include emotion, textual specificity, and decorative use of symbols. The emotion and specificity labels were assigned using external classifiers, as described in Appendix F. For decorative symbols, the presence of brackets was included, as they are widely used in Japanese ad texts to emphasize key information.

Analysis Method To analyze the relationship between each linguistic feature and human preference, we used the chi-square test of independence. This method assesses the independence between two categorical variables: (1) ad texts preferred by most evaluators and (2) superiority or inferiority of each linguistic feature. For example, when studying PPL, the relationship between preferred ad texts and their perplexity scores is analyzed.

Dataset We used 3,570 ad text pairs for which at least eight out of ten evaluators expressed a pref-

[8] Only a subset of features is presented in Table 4 due to space limitations. The complete list of 26 features, along with their definitions and analysis results, is in Appendix F.

	Linguistic Features	df	N	χ^2	ϕ
Raw text features	*Text length*				
	character$^{\dagger,\ddagger,\uparrow,*}$	1	2,925	723.8	0.497
Lexical features	*Content words*				
	noun$^{\dagger,\ddagger,\uparrow,*}$	1	1,406	326.6	0.482
	verb$^{\ddagger,\downarrow,*}$	1	535	6.9	0.114
	adjective	1	99	0.9	0.094
	Lexical choice				
	common noun$^{\dagger,\ddagger,\uparrow,*}$	1	1,397	288.1	0.454
	proper noun$^{\ddagger,\uparrow,*}$	1	152	7.6	0.223
	Character type				
	hiragana$^{\ddagger,\downarrow,*}$	1	2,047	23.2	0.107
	kanji$^{\ddagger,\uparrow,*}$	1	1,503	257.7	0.414
Syntactic features	*Dependency tree*				
	depth$^{\ddagger,\downarrow,*}$	1	1,914	16.9	0.094
	length	1	2,349	1.9	0.028
	Others				
	noun phrases$^{\dagger,\ddagger,\uparrow,*}$	1	1,895	259.8	0.370
	perplexity$^{\dagger,\ddagger,\downarrow,*}$	1	3,570	223.3	0.250
Stylistic features	*Emotion*				
	joy$^{\ddagger,\downarrow,*}$	1	693	70.1	0.318
	anticipation$^{\ddagger,\uparrow,*}$	1	683	89.3	0.362
	Others				
	specificity$^{\ddagger,\uparrow,*}$	1	186	116.4	0.791
	brackets$^{\dagger,\ddagger,\uparrow,*}$	1	1,667	1,372.6	0.907

Table 4: Results of the chi-square test. Df, N, and ϕ refer to the degree of freedom, the number of cases for each feature, and the measure of effect size, respectively. ‡ indicates linguistic features, identified in v2.0, that influence preference judgments, while † denotes those identified in v1.0. ↑ and ↓ indicate that ad texts with higher and lower feature scores, respectively, are preferred. $*$ indicates a significant relationship with human preferences ($p < 0.01$).

erence (§3.4), ensuring the reliability of the factor analysis influencing preferences. In addition, to focus on the differences in linguistic expressions between ad text pairs, we analyzed only the pairs with different scores for linguistic features, such that the number of cases for each feature varied. For example, 2,925 pairs had different character counts.

5.1.2 Results

Table 4 presents the chi-square test results. Linguistic features with higher chi-square values (χ^2) and lower p-values indicate a stronger relationship with human preferences. We also report Phi (ϕ), a commonly used measure of effect size for the chi-square test (Cohen, 1988). ϕ is defined as $\sqrt{(\chi^2/N)}$, where N is the number of observations. A value of 0.1 is considered a small effect, 0.3 a medium effect, and 0.5 a large effect.

These results reveal that several linguistic features, such as textual specificity and certain emotions (e.g., joy, anticipation), which were not identified by v1.0, are significantly related to human preferences. Specifically, cross-tabulations between linguistic features and preference judgments showed that ad texts with the following characteristics were preferred: *longer text*, *more nouns*, *shallower dependency trees*, *lower perplexity*, *higher specificity*, and the *presence of brackets*. These are examples of preferred features, and the full results are presented in Appendix F. Conversely, no significant differences were observed for features such as the number of adjectives.

5.2 Ad Text Generation

In this experiment, we focus on ad text refinement (Mishra et al., 2020), a task that generates more attractive ad texts by rephrasing the linguistic expressions without adding or removing any information.

5.2.1 Experimental Settings

Comparison Methods In exploring multiple methods for generating more attractive ad texts, we focused on recent LLM-based techniques, such as instruction tuning (Wei et al., 2022), preference tuning (Rafailov et al., 2023), and ICL (Brown et al., 2020). For ICL, we tested three types of prompts: (1) `zeroshot`, which provides only basic instructions for rephrasing an input ad text into a more attractive ad text; (2) `zeroshot-findings`, which further incorporates feature analysis findings (in §5.1) into the prompt; and (3) `fewshot-findings`, which extends `zeroshot-findings` by including 20 paraphrase examples sampled from the training data. As the findings, we incorporated higher character counts, greater fluency, and the use of brackets into the prompt. The few-shot examples were selected based on preference judgments, pairing less-preferred input texts with their corresponding preferred outputs. For instruction tuning, the LLMs were fine-tuned using less-preferred ad texts as inputs and highly preferred ad texts as outputs, based on human preference judgments. The instruction-tuned models were further refined by preference tuning via direct preference optimization (DPO) (Rafailov et al., 2023). For further implementation details, including the training setups and prompts used for each model, please refer to Appendix G.

LLMs Three LLMs, CALM3-22B (Ishigami, 2024), Swallow70B (Fujii et al., 2024), and GPT-4o (OpenAI, 2024), were employed. The first two

Model	PI	Att	Att&Length
CALM3-22B			
zeroshot	74.0	23.0	12.8
zeroshot-findings	74.0	42.6	23.0
fewshot-findings	85.0	38.8	**31.2**
instruct-zeroshot	**90.5**	31.5	29.3
dpo-zeroshot	70.5	**84.4**	8.5
Swallow70B			
zeroshot	90.5	15.5	8.3
zeroshot-findings	80.0	44.4	17.5
fewshot-findings	86.5	40.5	**26.0**
instruct-zeroshot	**94.0**	18.6	17.6
dpo-zeroshot	62.5	**71.2**	8.0
GPT-4o			
zeroshot	86.0	12.8	12.8
zeroshot-findings	**95.5**	**39.3**	**34.6**
fewshot-findings	92.5	37.8	33.5
Crowdworker	89.1	23.9	22.3

Table 5: Human evaluation results of ATG experiments. The evaluation used three metrics: PI, Att, and Att&Length, denoting the pass rate for paraphrase identification, the pass rate for attractiveness judgment, and the pass rate for attractiveness when length constraints are also considered, respectively.

models were chosen because they were pre-trained on Japanese corpora, either from scratch or through continual learning. We used GPT-4o via the Azure OpenAI API, version `2024-09-01-preview`. Additionally, to compare the human performance with those of LLMs, the paraphrases created by crowdworkers were evaluated. Crowdworkers were instructed to create paraphrases from the given ad text based on the guidelines described in §3.2.

Dataset A revised version of ADPARAPHRASE v2.0 was used for model training.[9] Specifically, the triplets (x, y_1, y_2) were formed by pairing source ad text x and two paraphrases y_1 and y_2 generated by the different models in v2.0. Subsequently, preference judgments were conducted for y_1 and y_2 using the annotation process in §3.4, collecting responses from ten evaluators. As a result, we constructed a dataset of 8,721 triplets (x, y_1^{pref}, y_2^{pref}), where y_1^{pref} and y_2^{pref} denote preference-labeled paraphrases. The dataset was split into training, development, and test sets at a ratio of $9 : 0.5 : 0.5$.

Evaluation Methods The generated texts were evaluated using three criteria: (1) paraphrase identification, (2) attractiveness, and (3) attractive while satisfying length constraints. Criteria (1) and (2) were assessed using the human evaluations described in §3.3 and §3.4. For (1), the percentage of generated texts judged as paraphrases by the majority of evaluators was calculated. For (2), among the texts judged as paraphrases, we reported the percentage judged as attractive by the majority. This evaluates the ability to generate an ad text that is both a valid paraphrase and attractive. For (3), among the texts judged as paraphrases, the percentage judged as attractive and satisfying the length constraint of 30 characters was determined. As ad texts that exceed length constraints cannot be delivered in online advertising, this metric evaluates the practical capability of generating attractive ad texts within the length constraint.

[9] In AdParaphrase v2.0, preference judgments were conducted on (x, y). However, this data format is not suitable for preference tuning such as DPO. Thus, the triplets (x, y_1, y_2) were created, and preference data were collected for (y_1, y_2).

Model	Perplexity↓	#Char↑	Brackets↑
CALM3-22B			
zeroshot	155.6	27.5	5.0
zeroshot-findings	157.6	30.7	64.5
fewshot-findings	146.7	27.0	**69.0**
instruct-zeroshot	168.5	24.1	48.5
dpo-zeroshot	**92.2**	**42.3**	37.0
Crowdworker	264.3	23.8	45.8
Source ad texts	169.7	23.6	39.5

Table 6: Linguistic features of generated ad texts. #Char and Brackets denote the average number of characters per text and the proportion of generated texts that include the bracket symbol, respectively.

5.2.2 Results

The evaluation results are presented in Table 5. For paraphrasing, the instruction-tuned methods demonstrated better performance. In terms of attractiveness, DPO-based models performed best overall. Furthermore, `zeroshot-findings` and `fewshot-findings`, which incorporate the findings of linguistic feature analysis, generated more attractive texts than `zeroshot`. This demonstrates that the findings obtained from the analysis contributed to improving the attractiveness of the generated texts. When considering attractiveness in conjunction with length constraints, the `zeroshot-findings` outperformed DPO-based models. This is because DPO-based models generated many texts that failed the length constraint, thereby reducing their score in this comparison.

Table 6 presents the linguistic features of the generated texts, including PPL, character count, and the presence of brackets, which were the key fea-

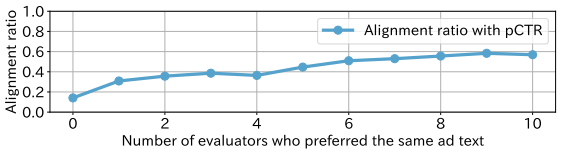

Figure 2: Alignment between human preferences and predicted Click-Through Rate (pCTR). The x-axis indicates human agreement level (number of evaluators with same preference). The y-axis shows the alignment ratio with pCTR. Higher human agreement correlates with increased alignment ratio, suggesting stronger consensus means better alignment.

tures incorporated into the prompt. The results indicate that models with higher attractiveness scores in Table 5 performed better across these linguistic features. Notably, DPO-based models exhibited higher character count. This suggests that DPO-based models tend to generate longer texts, potentially benefiting from length heuristics (Park et al., 2024), a bias where evaluators tend to perceive longer texts as more attractive.

6 Analysis

In this section, we report on the analyses conducted from two main perspectives, to contribute to the future development of attractive ad text generation. The first is an analysis of the relationship between human preferences and ad performance. Given that the ultimate goal of advertising is to optimize ad performance (e.g., clicks), clarifying the relationship between ad text preferences and ad performance is crucial. The second perspective concerns automatic evaluation of PI and attractiveness. Although PI and attractiveness were evaluated manually in this study, verifying automatic evaluation metrics as alternatives to manual evaluation is required to enhance efficiency in future research.

For the former perspective, we conducted two experiments: evaluating the relationship between human preference and predicted CTR (pCTR) (§6.1) and assessing ad performance in a real-world environment online (§6.2). For the latter, a meta-evaluation was performed to assess the relationship between human evaluation and existing automatic evaluation metrics (§6.3).

Ad delivery period		CTR↑	CVR↑	CTVR↑	CPC↓	CPA↓
Fitness	2 weeks	**91.5**	**141.4**	**129.4**	110.7	79.4
Education	2 weeks	77.7	249.0	200.0	100.5	40.4
	1 month	93.4	138.9	127.3	90.6	65.3

Table 7: Relative improvement of advertising performance metrics for different ad types (Fitness, Education) and delivery periods, compared to a baseline (100%). Bold values indicate statistically significant differences, as determined by a z-test ($p < 0.01$).

6.1 Relationship between Human Preferences and pCTR

It is critical to understand how the attractiveness of ad texts influences user behavior because the goal of advertising is to capture attentions and drive actions such as clicks. To explore this, we analyzed the relationship between human preferences and ad performance. Specifically, we examined whether the ad texts preferred by most evaluators also achieved a higher pCTR, a proxy for CTR.

Figure 2 shows the alignment between human preference and pCTR in AdParaphrase v2.0. The pCTR for each ad text was obtained using an in-house CTR prediction model. The x-axis represents the number of evaluators who preferred the same ad text, whereas the y-axis denotes the percentage of cases with pCTR and human preferences in agreement. For example, an x-axis value of ten means all evaluators preferred the same ad text in a pair, and the corresponding y-axis value shows the percentage of cases which also have higher pCTR. The results revealed a strong correlation between human preferences and pCTR (Pearson's correlation coefficient: 0.946), confirming that the ad texts preferred by the majority achieved higher CTRs. However, even when all the evaluators agreed on their preferences, the percentage of cases with a higher pCTR was approximately 60%, suggesting a potential upper limit for improving ad performance.

6.2 Online Evaluation of Ad Performance

In the online evaluation, we analyzed whether rephrasing ad texts into more attractive expressions influences ad performance, such as CTR. Specifically, we conducted an A/B test, comparing an existing group of ad texts with paraphrased ads generated using the `fewshot-findings` method[10] in §5.2. The tests were conducted on Google Ads, focusing on the headline text for ads from two com-

[10]For this evaluation, we used GPT-4 as the model.

Metrics	Paraphrase			Attractiveness		
	r	ρ	τ	r	ρ	τ
BLEU	**0.948**	0.950	0.831	-0.707	-0.484	-0.410
ROUGE-1	0.279	0.277	0.199	0.138	0.204	0.065
ROUGE-2	0.162	0.275	0.167	0.061	-0.113	-0.110
ROUGE-L	0.239	0.306	0.260	0.197	0.159	0.051
BERTScore	0.927	0.934	0.805	-0.769	-0.511	-0.385
GPT-4o	**0.948**	**0.965**	**0.895**	**0.886**	**0.758**	**0.615**

Table 8: System-level meta-evaluation results with Pearson (r), Spearman (ρ), and Kendall (τ).

panies in the fitness and education industries. The ads for the former ran for two weeks, whereas those for the latter ran for two weeks or one month. Details of the evaluation setup are provided in Appendix H.

Table 7 summarizes the relative improvement rates of paraphrased ads over existing ads using metrics such as CTR, conversion rate (CVR), CTVR, cost per click (CPC), and cost per action (CPA).[11] Among these, CTVR, defined as the product of CTR and CVR, is a comprehensive indicator of ad performance. The results indicate that as CTR decreases, CVR improves, reflecting actions such as purchases or sign-ups. Notably, for fitness ads, relative improvements in CVR and CTVR were statistically significant compared to the baseline.

6.3 Reliability of Automatic Evaluation Metrics

Adopting automatic evaluation methods is essential for enhancing efficiency in future studies. Thus, we analyzed whether existing automatic evaluation metrics can substitute human evaluations by conducting a system-level meta-evaluation. Specifically, we examined the correlations between human evaluation results from the ATG experiments (§5.2) and various automatic metrics. The evaluation metrics are presented in Table 8. Inspired by the LLM-as-a-judge paradigm (Gu et al., 2025), we included LLM-based evaluations using GPT-4o. For GPT-4o, we used human evaluation guidelines for PI and preference judgment as prompts. The LLM-based evaluation was reference-free, whereas the other metrics were reference-based, using human-created paraphrases (§5.2) as reference texts. The automatic evaluation scores are provided in Appendix G.

Table 8 presents the correlations between the scores and human evaluation results. For PI, BLEU, BERTScore, and GPT-4o exhibited strong positive correlations with human evaluations. With regard to attractiveness, GPT-4o showed a strong positive correlation, whereas BLEU and BERTScore displayed negative correlations. These results suggest that both reference-based and reference-free metrics are effective in predicting PI. However, reference-free metrics are more suitable for assessing attractiveness.

7 Conclusion

This study introduced AdParaphrase v2.0, a dataset for ad text paraphrasing that contains human preference data. Compared to v1.0, our dataset is 20 times larger, enabling a comprehensive analysis of the key features that make ad text attractive. We identified multiple linguistic features that contribute to engaging ad texts and investigated various methods for generating attractive ad texts. Our analysis revealed the relationship between human preference and ad performance, and demonstrated the potential of reference-free evaluation for assessing ad text attractiveness.

Future work will include enhancing ATG methods by addressing challenges such as adhering to length constraints, optimizing both human preference and ad performance, and investigating the influence of other factors on preferences, such as demographic information and product category.

8 Limitations

This study has several limitations that should be addressed in future studies.

Many Paraphrased Texts are LLM-Generated Many paraphrased texts are generated by LLMs, potentially resulting in linguistic features that differ from real ad texts. However, please note that the original CAMERA ads, used as source ad texts, were actually distributed ads, and so not all texts are LLM-generated. Future research could examine expression differences between human-written and LLM-generated ads or analyze how linguistic features influence preferences, focusing on human-authored texts.

Language-Specific Features and Generalizability AdParaphrase v2.0 is based on Japanese ad texts, meaning its linguistic feature analysis includes characteristics specific to Japanese, such as character types. However, other languages, such

[11] For advertising terms, see https://support.google.com/google-ads/topic/3121777.

as English and Chinese, also have unique linguistic features that may influence preferences, such as uppercase usage in English. It is important to note that our findings do not necessarily generalize to other languages. Future work could extend the dataset to multiple languages to explore whether certain linguistic features affecting preferences are shared across languages. To realize this multilingual extension, there are two possible approaches for multilingual adaptation: translating existing datasets like ADPARAPHRASE V2.0 or constructing new ones from scratch. Given that ads often include language- and region-specific proper nouns (e.g., product or service names), translation may lead to unnatural results. Therefore, we believe building datasets from scratch is more appropriate. This would involve collecting ad texts in the target language and applying the same process: paraphrase generation, identification, and preference annotation.

Limited Participants in Preference Judgments Due to time and financial constraints, the preference judgments were conducted with ten participants. Therefore, their preferences may not accurately reflect those of a broader population. To obtain more reliable and robust preference judgment results, collecting opinions from a larger number of participants is necessary. Additionally, this study recruited only Japanese participants. Since preferences can be influenced by demographic factors such as nationality, age, and gender, by collecting such additional information, it would be possible to analyze whether these factors influence preferences. An analysis incorporating demographic information would be a valuable future direction.

References

Kevin Bartz, Cory Barr, and Adil Aijaz. 2008. Natural language generation for sponsored-search advertisements. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 1–9.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems 33*, volume 33, pages 1877–1901.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1889–1905.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36*.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

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 and Data Mining*, pages 2269–2277.

Ryosuke Ishigami. 2024. cyberagent/calm3-22b-chat. Hugging Face.

Aditi Jha, Sam Havens, Jeremy Dohmann, Alex Trott, and Jacob Portes. 2023. Limit: Less is more for instruction tuning across evaluation paradigms. *arXiv preprint arXiv:2311.13133*.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104.

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.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop: Text Summarization Branches Out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 1483–1486.

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. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 2653–2660.

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 62th Annual Meeting of the Association for Computational Linguistics*, pages 955–972.

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.

Soichiro Murakami, Peinan Zhang, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2025. AdParaphrase: Paraphrase dataset for analyzing linguistic features toward generating attractive ad texts. *Preprint*, arXiv:2502.04674.

OpenAI. 2024. Hello gpt-4o. Accessed: 2025-01-03.

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.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017.

Reid Pryzant, Sugato Basu, and Kazoo Sone. 2018. Interpretable neural architectures for attributing an ad's performance to its writing style. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 125–135.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic assessment of Japanese text readability based on a textbook corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Victor O. Schwab. 2013. *How to Write a Good Advertisement: A Short Course in Copywriting*, illustrated edition edition. Echo Point Books & Media.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431.

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. 2018. Sudachi: a japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Hsueh-Cheng Wang and Marc Pomplun. 2012. The attraction of visual attention to texts in real-world scenes. *Journal of Vision*, 12(6):26–26.

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

Xiting Wang, Xinwei Gu, Jie Cao, Zihua Zhao, Yulan Yan, Bhuvan Middha, and Xing Xie. 2021. Reinforcing pretrained models for generating attractive text advertisements. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3697–3707.

Jason Wei, Maarten Paul Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew Mingbo Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations*.

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, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.

Brit Youngmann, Elad Yom-Tov, Ran Gilad-Bachrach, and Danny Karmon. 2020. The automated copywriter: Algorithmic rephrasing of health-related advertisements to improve their performance. In *Proceedings of The Web Conference 2020*, pages 1366–1377.

Yuan Yuan, Fengli Xu, Hancheng Cao, Guozhen Zhang, Pan Hui, Yong Li, and Depeng Jin. 2023. Persuade to click: Context-aware persuasion model for online textual advertisement. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1938–1951.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *The Eighth International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1298–1308.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1097–1100.

You are a professional copywriter responsible for creating search engine ads. Please rephrase the provided ad text to make it more attractive according to the following conditions

Conditions
- An ad text must be within 30 characters.
- Do not add new information or remove existing information from a given ad text.

Examples (30 paraphrase examples created by ad writers)
Input: Recommended in Kannai: Nail Salon
Output: Recommended Nail Salon @ Kannai

Input: Sell Gift Cards at a High Exchange Rate
Output: Sell Gift Cards at the Best Exchange Rate

Input: Up to ¥21,000 Discount for Online Applications
Output: Online Application: Up to ¥21,000 Discount
⋮

Answer
Additional conditions: **{stylistic instruction}**
Input: **{source ad text}**
Output: **{paraphrased ad text}**

Figure 3: Prompt for paraphrase candidate generation using LLMs.

Please rephrase the provided ad text to make it more attractive according to the following conditions.

Conditions
- An ad text must be within 30 characters.
- Do not add new information or remove existing information from a given ad text.

Tips for paraphrasing
Here are some tips for paraphrasing. However, you don't have to strictly follow them.
- Change the word order to make it clearer.
- Use simpler words.
- Choose more catchy expressions.
- Use synonyms with the same meaning.
- Add decorative symbols.
- Modify the character types.
- Place the most important information at the beginning.
- Use more abstract expressions.
- Use more specific expressions.
- Include words that encourage action.
- Use casual language.
- Turn statements into questions.

Examples
Input: Recommended in Kannai: Nail Salon
Output: Recommended Nail Salon @ Kannai

Input: Sell Gift Cards at a High Exchange Rate
Output: Sell Gift Cards at the Best Exchange Rate

Input: Up to ¥21,000 Discount for Online Applications
Output: Online Application: Up to ¥21,000 Discount
⋮

Answer
Input: **{source ad text}**
Output: **{paraphrased ad text}**

Figure 4: Guidelines for paraphrase candidate creation presented to crowd workers.

A Collecting Paraphrase Candidates

AdParaphrase v2.0 was constructed based on v1.0 and CAMERA, a Japanese ad text dataset. Both are governed by the CC BY-NC-SA 4.0 license, and we adhered to the intended use. The details of paraphrase candidate generation using LLMs and crowdworkers are as follows:

LLMs The prompt used to generate the paraphrase candidates is shown in Figure 3. For the few-shot examples, we used 30 paraphrase examples created by professional ad writers from AdParaphrase v1.0. In addition, to enhance paraphrase diversity, we defined 40 types of stylistic instructions, which are listed in Table 9. These instructions were defined based on previous studies on ATG (Kamigaito et al., 2021) and best practices in copywriting (Schwab, 2013). During paraphrase generation, a stylistic instruction was randomly selected for each source text. The effectiveness of these stylistic instructions is discussed in Appendix E. For all models, the temperature and top-p were set to 0.8 and 0.95, respectively.

Crowdworkers Figure 4 shows the annotation guidelines presented to the workers. The workers were given the same instructions and paraphrasing examples as those provided to the LLMs. To avoid increasing the annotation burden, we avoided providing explicit stylistic instructions to the human annotators, unlike the method used for LLMs. Because most workers had no prior experience in ad text creation, the guidelines also included tips on effective paraphrasing. These guidelines were developed based on insights from previous work (Kamigaito et al., 2021) on ATG and best practices in ad text creation (Schwab, 2013). We used Yahoo! Crowdsourcing as the crowdsourcing platform.[12] Native Japanese speakers were involved in the annotation process. Additionally, in accordance with the regulations of the crowdsourcing platform, each worker was compensated with 10 yen per task. The workers were informed in advance that their annotation results would be used for research purposes. Personally identifiable information was not obtained.

[12] https://crowdsourcing.yahoo.co.jp/

#	Instructions	#	Instructions
(1)	*Use many hiragana characters.*	(21)	*Use content words.*
(2)	*Use many katakana characters.*	(22)	*Use common words.*
(3)	*Use many kanji characters.*	(23)	*Use technical terms.*
(4)	*Write like a news article headline.*	(24)	*Use positive words.*
(5)	*Use more specific expressions.*	(25)	*Use negative words.*
(6)	*Use more abstract expressions.*	(26)	*Use neutral words.*
(7)	*Use first-person pronouns.*	(27)	*Use formal language.*
(8)	*Use second-person pronouns.*	(28)	*Use casual language.*
(9)	*Use third-person pronouns.*	(29)	*Place important information at the beginning.*
(10)	*Use expressions that convey excitement.*	(30)	*Place important information an the end.*
(11)	*Use expressions that convey joy.*	(31)	*Use more complex syntax.*
(12)	*Use calming and soothing expressions.*	(32)	*Use simpler syntax.*
(13)	*Use expressions that convey urgency.*	(33)	*Make it a question.*
(14)	*Use expressions that encourage action.*	(34)	*Use simple words.*
(15)	*Use brackets.*	(35)	*Use difficult words.*
(16)	*Use numbers.*	(36)	*Emphasize the benefits.*
(17)	*Use verbs.*	(37)	*Show how to solve the problem.*
(18)	*Use adjectives.*	(38)	*Include a catchy phrase.*
(19)	*Use nouns.*	(39)	*Use a visually clear expression.*
(20)	*Use adverbs.*	(40)	*Use an easy-to-read expression.*

Table 9: List of stylistic instructions for paraphrase candidate generation using LLMs

Please determine whether the following pair of ad texts is a paraphrase.

Criteria
Does the pair have the same meaning?
Does the pair convey the same content in different expressions?

Choices
- **{ad text #1}**
- **{ad text #2}**

Figure 5: Guidelines for paraphrase identification presented to crowd workers.

B Paraphrase Identification

Figure 5 presents the annotation guidelines for paraphrase identification provided to the workers. To ensure consistency between ADPARAPHRASE V1.0 and V2.0, we adopted the annotation guidelines used by Murakami et al. (2025). The criterion for paraphrase identification is whether two sentences convey the same meaning at the sentence level. We used Yahoo! Crowdsourcing as the crowdsourcing platform. The annotation workers were native Japanese speakers. Each worker was compensated with 10 yen per task in accordance with the regulations of the crowdsourcing platform. No personally identifiable information was collected during the annotation process. The workers were informed that their annotation results would be used for research purposes.

C Human Preference Judgments

Figure 6 presents the annotation guidelines for the preference judgments provided to the workers. To ensure that preference judgment criteria were consistent with ADPARAPHRASE V1.0, we adopted the same annotation guidelines as those used by Murakami et al. (2025). We used Yahoo! Crowdsourcing as the crowdsourcing platform. Native Japanese speakers were involved in the annotation process. In accordance with the regulations of the crowdsourcing platform, each worker was compensated with 10 yen per task. Personally identifiable information was not obtained. The workers were informed in advance that their annotation results would be used for research purposes.

D Example Paraphrase Pairs

Table 10 presents the example paraphrase pairs included in ADPARAPHRASE V2.0.

E Effect of Stylistic Instructions

We analyzed the effect of stylistic instructions on Principle (2). This analysis evaluated diversity from two perspectives: (1) differences between input and generated texts, and (2) diversity of generated texts. The first perspective uses a similarity score based on the Levenshtein edit distance (Levenshtein, 1966). The second perspective uses self-BLEU (Zhu et al., 2018). For self-BLEU, we measured lexical and syntactic diversity by evaluating word and part-of-speech (POS) sequences,

No.	Ad Text 1	Ad Text 2
(1)	*Canterbury/Official Online Store*	*Official Online Store of Canterbury*
(2)	*Recommended Hair Salon in Tama Center*	*Top Hair Salons – Tama Center*
(3)	*[2022] In-Depth Comparison of No-Installation WiFi*	*2022 Edition: Complete Comparison of No-Installation WiFi*
(4)	*Cheap Hotel Reservations in Kumamoto*	*Hotel Reservations in Kumamoto [Affordable]*
(5)	*If You're Serious About Solving Hair Problems*	*Hair Problems? Solve Them Seriously*
(6)	*[Official] Bleu : Blanc*	*Official Bleu Blanc Website*
(7)	*Budget Hotels Near Saiki Station*	*Book Cheap Hotels Near Saiki Station Now*
(8)	*[Official] Sumitomo Forestry Detached Homes for Sale*	*[Official] Sumitomo Forestry Residential Homes for Sale*
(9)	*Car Appraisal – Get the Best Price in 32 Seconds*	*What's the Highest Car Appraisal in 32 Seconds?*
(10)	*[Fine Save] Official Website*	*[Official] Fine Save Site*

Table 10: Example paraphrase pairs in ADPARAPHRASE V2.0.

Please select more attractive ad text between the two ad texts. Here are some criteria you can consider.

Criteria
- Which do you want to click on?
- Which is more memorable?
- Which is more attractive?
- Which is more eye-catching?
- Which is easier to understand?
- Which is easier to read?

Notes
If the impression of both ads is the same, please select "skip".

Choices
- {ad text #1}
- {ad text #2}
- skip

Figure 6: Guidelines for preference judgments presented to crowd workers.

	Levenshtein (↓) edit distance	Self-BLEU (↓) Word	Self-BLEU (↓) POS
Baseline prompt	0.542	27.8	90.3
w/ stylistic instruction	**0.504**	**26.4**	**88.9**

Table 11: Effects of stylistic instructions on diversity of paraphrase candidates generated by LLMs.

with POS diversity serving as a proxy for syntactic variation (Shaib et al., 2024). The results are summarized in Table 11, where lower scores for both perspectives indicate greater diversity. Introducing stylistic instructions improved the diversity of the paraphrase candidates generated by LLMs for both perspectives. These findings suggest that explicitly specifying textual styles in prompts effectively enhances the diversity of the generated texts.

Additionally, we analyzed the impact of each stylistic instruction on textual diversity. The results are summarized in Table 12, where each instruction number corresponds to an item in the stylistic instructions listed in Table 9. Our results show that the effects vary according to the instruction type. For example, *"Emphasize the benefits"* and *"Place important information at the end"* improved edit distance, whereas *"Include a catchy phrase"* enhanced lexical diversity and *"Use simpler syntax"* contributed to syntactic diversity.

F Analysis of Linguistic Features

F.1 Linguistic Features

Table 13 presents the 26 linguistic feature types used to analyze the factors influencing preference judgments. The definitions and extraction methods are described below. The extraction methods for each feature followed the same procedure outlined in Murakami et al. (2025).

Raw Text Features Character and word counts were used as raw text features because they affect the informativeness and readability of the text. Sudachi (Takaoka et al., 2018) was employed as a tokenizer for Japanese text.

Lexical Features The lexical features included the number of content words, character types, and lexical choices. Content words are indicative of the informativeness of ad texts, whereas character types are associated with readability (Sato et al., 2008). The number of content words was counted along with each character type. For the lexical choice, assuming that more frequently used words are preferred, the average word frequency was calculated using a balanced corpus of contemporary written Japanese (BCCWJ) (Maekawa et al., 2010). Additionally, the number of common and proper nouns was counted.

Syntactic features Syntactic features are measures of the complexity and fluency of ad texts. They include dependency tree depth, length of dependency links, number of noun phrases, and

#	Stylistic instruction	Levenshtein (↓) edit distance	Self-BLEU (↓) Word	Self-BLEU (↓) POS
(1)	Hiragana	0.472	29.6	83.4
(2)	Katakana	0.553	32.0	85.7
(3)	Kanji	0.577	31.6	87.2
(4)	News title	0.503	34.1	92.1
(5)	Specificity	0.524	35.4	92.5
(6)	Abstractness	0.368	27.4	89.1
(7)	First personal pronoun	0.527	36.1	90.8
(8)	Second personal pronoun	0.450	28.9	86.8
(9)	Third personal pronoun	0.537	34.2	88.5
(10)	Excitement	0.445	23.0	87.7
(11)	Joy	0.430	27.8	90.4
(12)	Ease	0.495	38.1	93.5
(13)	Urgency	0.485	31.1	92.2
(14)	Action	0.423	28.3	90.9
(15)	Brackets	0.622	44.6	92.0
(16)	Numbers	0.508	33.0	90.8
(17)	Verbs	0.523	39.6	90.9
(18)	Adjectives	0.552	35.6	88.9
(19)	Nouns	0.580	34.3	83.2
(20)	Adverbs	0.497	32.4	89.1
(21)	Content words	0.550	31.1	85.9
(22)	Common words	0.544	35.6	85.5
(23)	Uncommon words	0.499	28.1	85.9
(24)	Positive words	0.478	32.3	89.1
(25)	Negative words	0.479	26.2	86.8
(26)	Neutral words	0.520	34.9	83.0
(27)	Formal words	0.534	28.8	83.2
(28)	Casual words	0.438	31.6	90.1
(29)	Important words to left	0.465	40.1	88.6
(30)	Important words to right	0.382	41.5	88.7
(31)	Complex syntax	0.500	28.6	88.2
(32)	Simple syntax	0.549	32.5	**78.3**
(33)	Question	0.524	42.1	91.4
(34)	Simple words	0.498	35.4	83.7
(35)	Complex words	0.456	24.3	89.0
(36)	Benefits	**0.353**	23.9	92.5
(37)	Solutions	0.435	26.3	91.4
(38)	Catch-copy	0.392	**17.7**	87.3
(39)	Visibility	0.547	35.2	88.0
(40)	Readability	0.570	39.2	86.3

Table 12: Impact of each stylistic instruction on textual diversity. For each metric, the stylistic instruction exhibiting the largest impact is indicated in bold.

perplexity. Dependency parsing and noun phrase extraction were performed using spaCy with GiNZA [13]. Perplexity was calculated using GPT-2 [14] trained on web-crawled and Wikipedia corpora. The depth of a dependency tree is the longest path from the root to the leaf node, whereas the length of a dependency link is the number of words between the syntactic head and its dependent.

Stylistic Features Stylistic features included emotion, textual specificity, and decorative use of symbols in the text. Following Murakami et al. (2025), we assigned emotion and textual-specificity labels to each ad text using external classifiers. In addition to previously studied emotions such as *joy* and *anticipation*, we investigated *sadness* and *surprise*. Details of the classifiers can be found in §F.2. Regarding the decorative use of symbols, features such as the presence of brackets and question marks were considered. Brackets were included as features because they are widely used in Japanese ad text to emphasize important information and improve readability. Although question marks have not been studied in previous work, they are frequently used in ad texts to attract people's attention;

[13] https://github.com/megagonlabs/ginza

[14] https://huggingface.co/rinna/japanese-gpt2-medium

	Features	df	N	χ^2	ϕ
Raw text features	*Text length*				
	character$^{\dagger,\ddagger,\uparrow,*}$	1	2,925	723.8	0.497
	word$^{\ddagger,\uparrow,*}$	1	2,725	678.4	0.499
Lexical features	*Content words*				
	noun$^{\dagger,\ddagger,\uparrow,*}$	1	1,406	326.6	0.482
	verb$^{\ddagger,\downarrow,*}$	1	535	6.9	0.114
	adjective	1	99	0.9	0.094
	adjectival verb$^{\ddagger,\uparrow,*}$	1	105	8.0	0.276
	adverb	1	127	0.7	0.073
	Lexical choice				
	word frequency$^{\ddagger,\downarrow,*}$	1	2,666	70.8	0.163
	common noun$^{\dagger,\ddagger,\uparrow,*}$	1	1,397	288.1	0.454
	proper noun$^{\ddagger,\uparrow,*}$	1	152	7.6	0.223
	Character type				
	hiragana$^{\ddagger,\downarrow,*}$	1	2,047	23.2	0.107
	katakana$^{\ddagger,\uparrow,*}$	1	601	42.6	0.266
	kanji$^{\ddagger,\uparrow,*}$	1	1,503	257.7	0.414
	symbol$^{\ddagger,\uparrow,*}$	1	2,332	795.9	0.584
	digits$^{\ddagger,\uparrow,*}$	1	66	21.0	0.564
Syntactic features	*Dependency tree*				
	depth$^{\ddagger,\downarrow,*}$	1	1,914	16.9	0.094
	length	1	2,349	1.9	0.028
	Others				
	noun phrases$^{\dagger,\ddagger,\uparrow,*}$	1	1,895	259.8	0.370
	perplexity$^{\dagger,\ddagger,\downarrow,*}$	1	3,570	223.3	0.250
Stylistic features	*Emotion*				
	joy$^{\ddagger,\downarrow,*}$	1	693	70.1	0.318
	anticipation$^{\ddagger,\uparrow,*}$	1	683	89.3	0.362
	sadness$^{\ddagger,\downarrow,*}$	1	17	7.2	0.653
	surprise	1	28	0.2	0.083
	Others				
	specificity$^{\ddagger,\uparrow,*}$	1	186	116.4	0.791
	brackets$^{\dagger,\ddagger,\uparrow,*}$	1	1,667	1,372.6	0.907
	question marks	1	78	1.9	0.158

Table 13: Results of the chi-square test for linguistic features. Df, N, and ϕ refer to the degree of freedom and the number of cases, and the measure of effect size, respectively. ‡ indicates linguistic features, identified in v2.0, that influence preference judgments, while † denotes those identified in v1.0. ↑ and ↓ indicate that ad texts with higher and lower feature scores, respectively, are preferred. $*$ indicates a significant relationship with human preferences ($p < 0.01$).

Longer Ad Texts		Ad1	Ad2
Preferred Ad Texts	Ad1	1,308	549
	Ad2	200	868

Table 14: Example of Cross-tabulation between human preferences and number of characters in ad texts.

thus, we introduced them in this study.

F.2 External Classifiers

The following classifiers were used to assign labels for emotion and textual-specificity to each ad text. The use and construction of the classifiers followed the same procedure as that of Murakami et al. (2025).

Emotion The LUKE model[15] (Yamada et al., 2020), trained on WRIME (Kajiwara et al., 2021), a Japanese emotion analysis dataset based on social media text, was used to label the emotions in ad texts. This model is an eight-class classifier that assigns the most appropriate emotion from the following eight categories: *joy, sadness, anticipation, surprise, anger, fear, disgust, and trust*. The classifier achieved an accuracy of 68.6%.

Textual Specificity A specificity classifier was created using GPT-4 via the Azure OpenAI API (`2024-09-01-preview`) with a few-shot setting. This task was formulated as a three-class classification problem, in which the model compared two ad texts to determine which has higher specificity. If both had equivalent specificity, a label of *"equal"* was output. To evaluate model performance, 100 predictions were randomly sampled and manually evaluated, achieving an accuracy of 88.0%.

F.3 Results

Table 13 presents the results of the chi-square test for all the features. Several features that were not identified in v1.0 (Murakami et al., 2025) were found to be strongly related to preference judgments. Specifically, ‡ indicates the linguistic features identified in v2.0that influenced preference judgments, whereas † denotes those identified in v1.0.

In addition, we analyzed the relationship between each feature and human preferences by cross-tabulating feature values. Table 14 shows the cross-tabulation of preference judgments and text length

[15] https://huggingface.co/Mizuiro-sakura/luke-japanese-large-sentiment-analysis-wrime

```
You are a professional copywriter responsible for creating
search engine ads. Please rephrase the provided ad text to
make it more attractive according to the following conditions.

# Conditions
- An ad text must be within 30 characters.
- Do not add new information or remove existing information
from a given ad text.

# Tips for Creating Attractive Ad Texts                Findings
- Utilize bracket symbols such as 【】 or 「」.
- Use the full character limit (30 characters).
- Make the text more fluent.

# Examples (20 cases)                          Few-shot examples
Input: 4-minute walk from Fujiidera Station
Output: From Fujiidera Station, a 4-minute walk

Input: Experienced Talent Hiring [For Corporations]
Output: Immediate Hiring of Experienced Talent for
Corporations
                    ⋮
# Answer
Input: {ad text}
Output: {paraphrased ad text}
```

Figure 7: Prompt for `fewshot-findings` in ATG experiment.

(in characters) for ad text pairs, where Ad 1 and Ad 2 refer to the source and paraphrased ad text, respectively. In Table 14, 1,308 cases were observed, where Ad 1, which has a higher character count, was preferred by the majority of evaluators. We performed this analysis for each feature. In Table 13, the ↑ and ↓ symbols indicate that ad texts with higher and lower feature scores, respectively, are preferred. For example, we found that ad texts with the following characteristics are preferred: *longer text*, *more nouns*, *lower dependency tree*, *lower perplexity*, and *inclusion of brackets*, suggesting that these features are key for enhancing the attractiveness of ad texts.

G Ad Text Generation

Implementation Details of ATG Models Figure 7 presents the prompt for the `fewshot-findings` model, which is an ICL-based approach. The few-shot examples consist of 20 paraphrases randomly sampled from the training data, and the findings incorporate insights from linguistic feature analysis, encouraging longer, more fluent sentences and use of brackets. The `zeroshot` model relies solely on basic instructions and excludes few-shot examples and findings, whereas the `zeroshot-findings` model incorporates only the findings. The instruction-tuned and DPO models were implemented using a quantized low-rank

Model	BL	R-1	R-2	R-L	BS	GPT-4o	
						Para	Att
CALM3-22B							
zeroshot	27.4	29.3	9.8	29.2	86.5	75.0	33.0
zeroshot-findings	30.2	30.8	10.5	31.0	86.8	66.5	49.0
fewshot-findings	40.0	32.0	10.5	31.3	89.5	79.0	25.5
instruct-zeroshot	**46.8**	**32.4**	**11.0**	**32.3**	**90.3**	**89.5**	11.0
dpo-zeroshot	15.9	29.8	10.5	29.7	81.5	59.5	**80.0**
Swallow-70B							
zeroshot	41.8	31.0	**11.5**	31.2	89.2	90.0	11.5
zeroshot-findings	37.8	31.5	10.3	31.0	87.8	78.0	32.5
fewshot-findings	44.5	**32.5**	10.5	**31.6**	89.2	83.0	18.0
instruct-zeroshot	**50.5**	29.4	10.7	29.0	**90.9**	**90.5**	6.5
dpo-zeroshot	20.1	30.0	10.5	29.7	82.8	61.5	**65.5**
GPT-4o							
zeroshot	37.7	27.1	9.2	26.9	88.1	90.0	3.0
zeroshot-findings	48.0	31.2	9.5	31.0	90.7	**92.0**	5.5
fewshot-findings	**49.5**	**32.3**	**10.9**	**31.5**	**91.0**	90.5	**7.5**

Table 15: Automatic evaluation results of ATG experiment.

Model	Perplexity↓	#Char↑	Brackets↑
CALM3-22B			
zeroshot	155.6	27.5	5.0
zeroshot-findings	157.6	30.7	64.5
fewshot-findings	146.7	27.0	**69.0**
instruct-zeroshot	168.5	24.1	48.5
dpo-zeroshot	**92.2**	**42.3**	37.0
Swallow70B			
zeroshot	158.3	27.7	13.0
zeroshot-findings	129.3	32.9	**89.0**
fewshot-findings	116.8	29.7	63.5
instruct-zeroshot	170.7	23.7	39.0
dpo-zeroshot	**70.5**	**42.4**	42.0
GPT-4o			
zeroshot	236.9	21.5	34.5
zeroshot-findings	228.9	25.0	**100.0**
fewshot-findings	**183.8**	**25.7**	73.0
Crowdworker	264.3	23.8	45.8
Source ad texts	169.7	23.6	39.5

Table 16: Linguistic features of generated ad texts.

adaptation (QLoRA) (Dettmers et al., 2023) and trained for one epoch. The implementation followed the code in the repository[16]. Greedy decoding was used during inference.

Automatic Evaluation Metrics For automatic evaluation, multiple metrics were used, including BLEU (BL) (Papineni et al., 2002), ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) (Lin, 2004), BERTScore (BS) (Zhang et al., 2020), and LLM-based evaluation with GPT-4o (Liu et al., 2023; Gu et al., 2025). These metrics were chosen for their widespread use in paraphrase and other text generation tasks. The F1 scores are reported for ROUGE and BERTScore. For LLM-based evaluation, we used human evaluation guidelines for PI and preference judgments as prompts. These guidelines are displayed in Figures 5 and 6. The LLM-based evaluation is reference-free, whereas the other metrics are reference-based. For reference-based metrics, the human-created paraphrases in §5.2 were used as the reference text.

Automatic Evaluation Results Table 15 presents the automatic evaluation results. Here, we report the results of a single run. In automatic evaluations using BLEU, ROUGE, BERTScore, and GPT-4o for PI, the instruction-tuned model and `fewshot-findings` outperformed the other models. Conversely, in the GPT-4o evaluation of attractiveness, the DPO model achieved the best performance.

[16] https://github.com/ghmagazine/llm-book

Linguistic Features of Generated Texts Table 16 presents the linguistic features of the generated texts for all models, including PPL, character count, and presence of brackets, which were the key features incorporated into the prompt. These results suggest that the models with higher attractiveness scores in Table 5 performed better across these linguistic features. Notably, DPO-based models exhibited lower PPL and greater character counts, indicating that these factors contribute to the attractiveness of the generated ad texts.

H Online Evaluation

In the online evaluation, ad texts paraphrased using the `few-shot-findings` method described in §5.2 were deployed to analyze whether paraphrasing more attractive expressions influenced user behavior, such as clicks. Specifically, an A/B test was conducted to compare an existing ad group as baseline with a paraphrased ad group. This evaluation was conducted using Google Ads. In search advertising, each ad consisted of 15 headlines and 3 descriptions. The paraphrasing method was applied to the headlines of the existing ads, whereas the descriptions remained the same as those in the baseline. Ads from two companies in the fitness and education industries were used for evaluation, and prior consent was obtained. The ads from the first company were deployed for two weeks. For the second company, ads were deployed twice for different durations. The results from the two-week and one-month deployments are reported.

Table 7 presents the results of the online evaluation, comparing the click-through rate (CTR), conversion rate (CVR), CTVR, cost per click (CPC), and cost per action (CPA) between the existing and paraphrased ads. Here, CTVR is the product of CTR and CVR and serves as a comprehensive metric for evaluating ad effectiveness. CPC and CPA represent the costs incurred per click and action, respectively; lower values are preferable for cost efficiency.

Statistical significance tests were conducted using the z-test for CTR, CVR, and CTVR. For each metric, the z-test compares the rate between the tested ad and baseline, thereby calculating a z-value based on the underlying counts to derive a p-value. The p-values below the significance level (0.01) indicated a statistically significant difference for that metric.