Striking Gold in Advertising: Standardization and Exploration of Ad Text Generation

Anonymous ACL submission

Abstract

In response to the limitations of manual ad creation, significant research has been conducted in the field of automatic ad text generation (ATG). However, the lack of comprehensive benchmarks and well-defined problem sets has made comparing different methods challenging. To tackle these challenges, we standardize the task of ATG and propose a first benchmark dataset, ATG-BENCH , carefully designed and enabling the utilization of multi-modal information and facilitating industry-wise evaluations. Our extensive experiments with a variety of nine baselines, from classical methods to state-of-the-art models including large language models (LLMs), show the current state and the remaining challenges. We also explore how existing metrics in ATG and an LLM-based evaluator align with human evaluations.

1 Introduction

The global online advertising market has witnessed significant growth and quadrupled over the last decade, particularly in the domain of search ads (Meeker and Wu, 2018) . Search ads are designed to accompany search engine results and are tailored to be relevant to users' queries (search queries) (Figure 1). These ads are displayed alongside a landing page (LP), providing further details about the advertised product or service. Therefore, ad creators must create compelling ad texts that captivate users and encourage them to visit the LP. However, the increasing volume of search queries, which is growing at a rate of approximately 8% annually (Djuraskovic, 2022), poses challenges for manual ad creation.

The growing demand in the industry has fueled research on the automatic generation of ad texts. Researchers have explored various approaches, starting with *template-based* methods that generate ad text by inserting relevant keywords into predefined templates (Bartz et al., 2008; Fujita et al., 2010; Thomaidou et al., 2013). Recently, neural language generation (NLG) techniques based on encoder-decoder models, which are widely employed in machine translation and automatic summarization, have been applied to ad text generation (ATG) (Hughes et al., 2019; Mishra et al., 2020; Kamigaito et al., 2021).

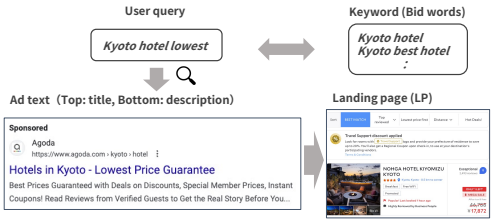


Figure 1: An example of search ads.

However, the automated evaluation of ATG models presents significant challenges. Previous research has been constrained to conducting individual experiments using proprietary datasets that are not publicly available (Murakami et al., 2023). This limitation arises from the absence of a shared dataset (i.e., a benchmark) that can be universally applied across the field. Moreover, the absence of benchmarks has resulted in a lack of consensus regarding task settings such as the models' input/output formats. While some studies use keywords as input (Bartz et al., 2008; Fukuda, 2019), others employ existing advertisements (Mishra et al., 2020) or LPs (Hughes et al., 2019; Kanungo et al., 2022; Golobokov et al., 2022). This variation in the task setting indicates that the field as a whole has yet to establish a standardized problem setting, which hinders the generalization and comparability of ATG techniques.

This study aims to advance ATG technology by standardizing the task setup, transforming it into a format accessible to potential players by providing a shared dataset and exploring the current status and limitations. Standardizing problem settings

common to a variety of advertising applications as tasks allows for focused exploration of core issues in an academic context while maintaining the flexibility to be applied to a wide variety of applications (§3). To engage a broader community of researchers beyond those who possess ad data, we construct the first publicly available benchmark, **ATG-BENCH** , which is meticulously developed a comprehensive dataset (§4). Our dataset comprises actual data sourced from Japanese search ads and incorporates annotations encompassing multimodal information such as the LP images. To explore the current state and future challenges, we conducted extensive experiments using nine diverse baselines, including multimodal models and large language models (LLMs), as well as the dominant approaches in existing studies (§5). Furthermore, we also conducted a meta-evaluation of how well the existing metrics and LLM-based evaluators reproduced human evaluations (§6).

Our major contributions are:

- Lowering entry barriers in ATG through task standardization, the creation of the initial benchmark, and public dataset sharing.[1]
- Benchmarking experiments with nine diverse models, including classical, standard, and state-of-the-art LLM-based models, demonstrated the current state and future challenges.
- The first meta-evaluation in ATG highlighted the reliability and limitations of commonly used metrics.

We observed the following:

- Fine-tuned encoder-decoder models play an important role in maximizing automatic evaluation scores and improving quality in intrinsic evaluations such as faithfulness and fluency.
- Few-shots with strong LLMs have great potential for quality improvement in extrinsic evaluations such as human preference.
- Using multimodal information like LP images improves ad quality, but methods for model integration require further exploration.
- Model performance and rankings vary by industry domain.
- Existing metrics work as intrinsic evaluations, but it is still difficult to use them as a substitute for extrinsic evaluations.

[1] https://github.com/anonymized

2 Background

Various types of online advertising exist, including search ads, display ads [2], and slogans [3]. However, since most existing studies are related to search ads (Murakami et al., 2023), this study also focuses on search ads and provides an overview of ATG research and its current limitations.

2.1 A quick retrospective

Early ATG systems predominantly relied on template-based approaches (Bartz et al., 2008; Fujita et al., 2010; Thomaidou et al., 2013). These approaches involved filling appropriate words (i.e., keywords) into predefined templates, resulting in the generation of ad texts. Although this method ensured grammatically correct ad texts, it has limitations in diversity and scalability because it could only accommodate variations determined by the number of templates, which are expensive to create. To address these constraints, alternative approaches have been explored, including reusing existing promotional text (Fujita et al., 2010) and extracting keywords from LPs to populate template slots (Thomaidou et al., 2013).

Encoder-decoder models, which have demonstrated their utility in NLG tasks such as machine translation and summarization (Sutskever et al., 2014), have been applied to ATG research (Hughes et al., 2019; Youngmann et al., 2020; Kamigaito et al., 2021; Golobokov et al., 2022). These models have been employed in various approaches, including *translating* low click-through-rate (CTR) sentences into high CTR sentences (Mishra et al., 2020), *summarizing* crucial information extracted from the LPs (Hughes et al., 2019; Kamigaito et al., 2021), and combining these techniques by first summarizing the LPs and subsequently translating them into more effective ad texts based on CTR (Youngmann et al., 2020).[4] Recently, transfer learning approaches using pre-trained language models have become mainstream, allowing for more fluent and

[2] Display ads typically take the form of banner ads strategically placed within designated advertising spaces on websites or applications.

[3] Slogans are catchy phrases designed to captivate the attention of internet users and generate interest in products, services, or campaigns.

[4] CTR is a widely-used indicator of advertising effectiveness in the online advertising domain.

Work	Approach	Input	Output	Affiliation	Lang.	xACL
Bartz et al. (2008)	Template	Keyword	Ad text	Yahoo	En	
Fujita et al. (2010)	Template	Promotional text	Ad text, Keyword	Recruit	Ja	
Thomaidou et al. (2013)	Template	LP	Ad text	Athens Univ.	En	
Hughes et al. (2019)	Seq2Seq	LP	Ad text	Microsoft	En	
Fukuda (2019)	Seq2Seq	Keyword	Ad text	DENTSU	Ja	
Mishra et al. (2020)	Seq2Seq	Ad text	Ad text	Yahoo	En	
Youngmann et al. (2020)	Seq2Seq	LP, Ad text	Ad text	Microsoft	En	
Duan et al. (2021)	Seq2Seq	Query, KB	Ad text	Tencent	Zh	
Kamigaito et al. (2021)	Seq2Seq	LP, Query, Keyword	Ad text	CyberAgent	Ja	✓
Wang et al. (2021)	Seq2Seq	LP, Ad text	Ad text	Microsoft	En	
Zhang et al. (2021)	Seq2Seq	Ad text, Keyword, KB	Ad text	Baidu	Zh	
Golobokov et al. (2022)	Seq2Seq	LP	Ad text	Microsoft	En	✓
Kanungo et al. (2022)	Seq2Seq	Multiple ad texts	Ad text	Amazon	En	
Wei et al. (2022)	Seq2Seq	User review, Control code	Ad text	Alibaba	Zh	✓
Li et al. (2022)	Seq2Seq	Query	Ad text, Keyword	Microsoft	En	✓
Murakami et al. (2022a)	Seq2Seq	Keyword, LP	Ad text	CyberAgent	Ja	

Table 1: A summary of existing research on ad text generation. *xACL* (✓) presents whether the paper belongs to the ACL community, or some other research community (no ✓).

diverse ATG (Wang et al., 2021; Zhang et al., 2021; Golobokov et al., 2022; Kanungo et al., 2022; Wei et al., 2022; Li et al., 2022; Murakami et al., 2022a).

2.2 Current limitations

ATG has experienced remarkable growth in recent years, garnering significant attention as a valuable application of natural language processing (NLP). However, the automated evaluation of models presents substantial challenges. Existing studies, validated only on *non-public* datasets, hinder fair comparisons and discussions across studies, posing challenges in generalizing ATG technology. Related to this, the problem settings for ATG, such as input/output, are not shared among the studies because there are variations depending on the advertising medium (e.g., search ads, display ads, etc.) and platform (Google, Bing, Yahoo, etc.). These challenges are primarily due to the absence of a shared benchmark dataset that can benefit the entire research community. The reason behind the reluctance to share ad datasets is that they usually contain performance values such as CTR, which are confidential data for companies. Table 1 summarizes the existing studies in the field and shows that this field is led by companies operating advertising-related businesses. ATG is gaining significant attention within the ACL community as a promising application of NLP. Moreover, it stands out as a valuable research subject contributing to the development of human-centered NLP techniques, as discussed in §3. As a confluence of these trends, this study aims to establish ATG as an NLP task by standardizing the task and building a benchmark dataset.

3 Standardization of ad text generation

One of the goals of this study is to develop a task that is not specific to a particular platform or advertising medium but focuses on universal core problems common to these applications, to facilitate generalization of ATG technology. To meet these requirements, we standardize the ATG task as follows: Let $\boldsymbol{x}$ be a source document that describes advertised products or services, $\boldsymbol{a}$ a user signal reflecting the user's latent needs or interests, and $\boldsymbol{y}$ an ad text. ATG aims is to model $\boldsymbol{p(y|a,x)}$. User signals, such as search keywords for search ads and user browsing and action history for display ads, can vary based on the application and domain. The specific data to be selected for each $\boldsymbol{x}$, $\boldsymbol{a}$, and $\boldsymbol{y}$ will be left to future dataset designers and providers. This standardization of ATG allows a focused exploration of core issues in an academic context while maintaining flexibility for diverse applications in an industrial context.

The requirements of ad text The purpose of advertising is to influence consumers' (users) attitudes and behaviors towards a particular product or service. Therefore, the goal of ATG is to create text that encourages users' purchasing behaviors. In this study, we have identified the following two fundamental requirements of ad text: (1) The information provided by the ad text is consistent with the content of the source document; and (2) the information is carefully curated and filtered based on the users' potential needs, considering the specific details of the merchandise. Requirement 1 relates to *hallucinations*, which is currently a highly prominent topic in the field of NLG (Wiseman et al., 2017; Parikh et al., 2020; Maynez et al., 2020). This requirement can be considered crucial

for practical implementation since the inclusion of *non-factual hallucination* in ad texts can cause business damage to advertisers. Regarding requirement 2, it is necessary to successfully convey the features and attractiveness of a product within a limited space and immediately capture the user's interest. Therefore, ad text must selectively include information from inputs that can appeal to users.

Differences from existing tasks The ATG task is closely related to the conventional document summarization task in that it performs information compression while maintaining consistency with the input document's content. Particularly, *query-focused summarization (QFS)* (Dang, 2005), a type of document summarization, is the closest in problem setting because it takes the user's query as the input; however, there are some differences. The task of QFS aims to create a summary from one or multiple document(s) that answers a specific query (*explicit needs*). In contrast, ATG is required to extract not only surface information from user signals but also the *latent needs* behind them and then return a summary. For example, when a user's query is "used cars," the goal of QFS is to provide information about used cars. On the other hand, for users seeking higher-priced items like cars, factors such as quality become important even if they are used. Therefore, the task of ATG aims to present ads that include expressions appealing to high quality and reassurance, such as *"All cars come with a free warranty!"*.

Another notable difference is that while summarization aims to deliver accurate text that fulfills task-specific requirements, ATG surpasses mere accuracy and aims to influence user attitudes and behavior. Consequently, unconventional and/or ungrammatical text may be intentionally used in ad-specific expressions to achieve this objective (refer to details in §4.2). Therefore, QFS is a subset of ATG (QFS $\subset$ ATG). One of the technical challenges unique to ATG is capturing users' latent needs based on such user signals $\boldsymbol{a}$ and generating appealing sentences that lead to advertising effectiveness, which depends significantly on the psychological characteristics of the recipient users. Therefore, realizing more advanced ATG will also require a connection with advertising psychology (Scott, 1903) based on cognitive and social psychology. The ATG is an excellent research topic for advancing user-centered NLP technologies.

	# instance	# ad text	Industry-wise
Train	12,395	1	
Dev	3,098	1	
Test	872	4	✓

Table 2: Statistics of our dataset. *Industry-wise* (✓) indicates whether the data is separable by industry.

4 Construction of ATG-BENCH

4.1 Dataset design

In this study, the following two benchmark design policies were first established: the benchmark should be able to (1) utilize multimodal information and (2) evaluate by industry domain. In terms of **Design Policy 1**, various advertising formats use textual and visual elements to communicate product features and appeal to users effectively. It is well-recognized that aligning content with visual information is crucial in capturing user attention and driving CTR. Exploring the effective utilization of such multi-modal information is crucial for the ATG. **Design Policy 2** highlights the significance of incorporating specific *advertising appeals* to create impactful ad texts. In general, ad creators must consider various aspects of advertising appeals such as the *price*, *product features*, and *quality*. For instance, advertising appeals in terms of *price* such as *"free shipping"* and *"get an extra 10% off"* captivate users by emphasizing cost savings through discounts and competitive prices. Previous studies revealed that the effectiveness of these advertising appeals varies depending on the target product and industry type (Murakami et al., 2022b). To foster the development of robust models, it is crucial to conduct an industry-wise evaluation.

4.2 Construction procedure

We utilized Japanese search ads from our company involved in the online advertising business.[5] In these source data, the components of user queries, ad texts, and LPs (URLs) are allocated accordingly. Search ads comprise a *title* and *description*, as shown in Figure 1. Description in search ads has a larger display area compared to titles. It is typically written in natural sentences but may also include advertising appeals. In contrast, titles in search ads often include unique wording specific to the advertisements. They may deliberately break or compress grammar to the extent acceptable to

[5]Careful care is taken to ensure that advertisers are not disadvantaged in the data release.

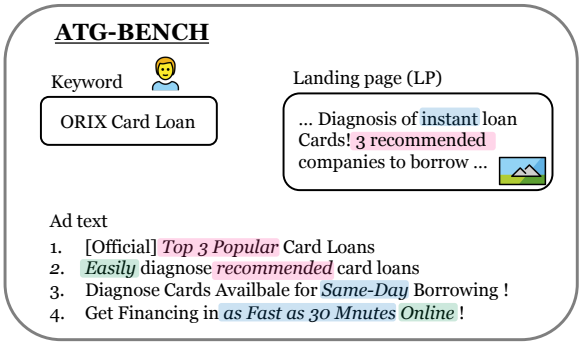


Figure 2: Examples of our dataset, translated into English for visibility. The highlighted areas in each color indicate the aspects of advertising appeals: *Speed*, *Trend*, and *User-friendliness*, based on Murakami et al. (2022b)'scheme.

humans because their primary role is immediately capturing a user's attention. For instance, the sentence *"If you're looking to sell your brand-name merchandise, why not get a free valuation at XX right now?"* is transformed into an ad-specific expression: *"Sell your brand-name goods / free valuation now"*. Studies in advertising psychology have reported that these seemingly ungrammatical expressions, unique to advertisements, not only do not hinder human comprehension but also capture their attention (Wang et al., 2013). We extracted only titles as ad texts $\boldsymbol{y}$ to create a benchmark focusing on ad-specific linguistic phenomena.

In our dataset, we extracted `meta description` from the HTML-associated LPs, which served as a description document (*LP description*) $\boldsymbol{x}$ for each product. Furthermore, in line with **Design Policy 1**, we processed a screenshot of the entire LP to obtain an LP image, allowing us to leverage multi-modal information. Through this process, we obtained images $\boldsymbol{I}$, layout information $\boldsymbol{C}$, and text $\{x_i^{\texttt{ocr}}\}_{i=1}^{|\boldsymbol{R}|}$ for the rectangular region set $\boldsymbol{R}$ using the OCR function of the Cloud Vision API.[6]

4.3 Annotation

The source data is assigned a delivered gold reference ad text, but because of the variety of appeals in the ads, there is a wide range of valid references for the same product or service. Therefore, three additional gold reference ad texts were created for the test set by three expert annotators who are native Japanese speakers with expertise in ad annotation. As explained in §3, since it is important for ad creation to consider latent needs behind user signals, we instructed the annotators to explicitly consider search keywords as user intentions.[7] During the data collection process for evaluation annotations, data were randomly selected based on keywords manually mapped to industry labels, such as *"designer jobs"* mapped to the human resource industry, following **Design Policy 2**. Here, we used the following four industry domain labels: human resources (HR), e-commerce (EC), finance (Fin), and education (Edu).

Table 2 provides the statistics of our dataset. The dataset was partitioned into training, development, and test sets to prevent data duplication between the training (development) and test sets, which was achieved through filtering processes. Figure 2 presents examples from the test set of this dataset.[8] Although the annotator was not given explicit instructions regarding the advertising appeal, we confirmed that the annotator created an ad text (#2-4) that featured a variety of advertising appeals different from the original ad text (#1) that considered latent needs based on keywords. This suggests that our test set captures a certain level of diversity in expressing advertisements.

5 Benchmarking of ATG models

To clarify the current state and remaining challenges, we conduct benchmark experiments using the dataset constructed in §4 and various ATG models. Specifically, we investigate the following research questions:

RQ1 *How do differences in the use of pre-trained language models (i.e., finetuning vs. few-shot) affect overall performance?*

RQ2 *Is multimodal information useful for ad text generation?*

RQ3 *Do trends in model performance vary by industry domain?*

RQ4 *What are the qualitative differences between generated ad text compared to human-produced ad text?*

5.1 Models

As outlined in §2.2, existing studies use non-public data with performance values, such as CTRs, and

[6] https://cloud.google.com/vision/docs/ocr

[7] The detailed annotation guidelines are presented in Appendix A.

[8] Although not included due to space limitations, the actual dataset also includes LP images (screenshots), their OCR results, and industrial labels.

therefore cannot be replicated on the ATG-BENCH data set, which does not include performance values. Therefore, this experiment will focus on a simplified replication of previous studies and follow-up on the dominant approach.

- **BM25** is a model of an extractive approach using the BM25 algorithm (Robertson et al., 2009). The BM25 algorithm is used to generate ad texts by extracting one query-related sentence from the input document.

- **BART** is a fine-tuned model using BART (Lewis et al., 2020) . We used the following pre-trained model: `japanese_bart_base_2.0` [9]

- **T5** is a fine-tuned model using T5 (Raffel et al., 2022). We used the following pre-trained model: `sonoisa/t5-base-japanese` [10].

- **GPT-3.5** is a few-shot model using GPT-3.5 (`gpt-3.5-turbo-0613`) (Ouyang et al., 2022). We built the model using the API provided by OpenAI [11].

- **GPT-4** is a few-shot model using GPT-4 (`gpt-4-0613`) (OpenAI, 2023). As with GPT-3.5, we constructed the model using the API provided by OpenAI.

- **Llama2** is a few-shot model using Llama2 (Touvron et al., 2023). We used the following pre-trained model: `ELYZA-japanese-Llama-2-7b-instruct` [12].

For BART and T5, we fine-tuned each pre-trained model on the train split of ATG-BENCH to create our baseline models. For GPT-3.5, GPT-4, and Llama2, the baseline models were constructed by 3-shot in-context learning, respectively. To investigate the effectiveness of incorporating multi-modal features such as images and layout in the LPs and their impact on the overall performance, we built various settings for the T5-based model that considered LP image information, following Murakami et al. (2022a). Specifically, we incorporated the following three types of multi-modal information into the model architecture: LP OCR text (`lp_ocr;o`), LP layout information (`lp_layout;l`), and LP BBox image features (`lp_visual;v`)[13]. See Appendix B for details on the experimental setup for each baseline model, including the prompt template.

5.2 Evaluation

Automatic evaluation To evaluate the generated texts quality, we employed two widely used metrics in ATG (Murakami et al., 2023): BLEU-4 (B-4)[14] (Papineni et al., 2002) and ROUGE-1 (R-1) (Lin, 2004). These metrics assess the similarity between the generated text and reference based on n-gram overlap. Since paraphrases are commonly used in ad texts, BERTScore (BS) (Zhang et al., 2020), an embedding-based metric, was also used to handle their semantic similarity. Additionally, as task-specific guardrails, we introduce keyword insertion rates (KWD) (Mishra et al., 2020) and sentence length regulation compliance rates (REG). KWD represents the percentage of cases where the specified keyword is included in the generated text for evaluating the relevance of the LP and the ad text. REG indicates the percentage of compliance with the character count regulation (15 characters or less).

Manual evaluation To answer RQ4, we conducted a manual evaluation. Three human raters evaluate each of the 10 ad texts of the 9 models (§5.1) and one original reference for each of the three evaluation aspects of *faithfulness*, *fluency*, and *attractiveness (i.e. human preference)*. The faithfulness and fluency evaluations were conducted using an *absolute* evaluation of whether the input document implies or does not imply the ad text, and whether the content of the ad text is understandable and natural, respectively. Given the challenge of providing an absolute evaluation of each ad text's attractiveness, we conducted a pairwise evaluation comparing the human reference and each model output, considering cases where the attractiveness was equal (*Tie*). For faithfulness and fluency, we sampled 200 cases from the test data and conducted manual evaluations for a total of 2000 ad texts. For attractiveness, we sampled 100 cases, created pairs of the human reference and each model output, and

[9] https://github.com/utanaka2000/fairseq/tree/japanese_bart_pretrained_model

[10] https://huggingface.co/sonoisa/t5-base-japanese

[11] https://github.com/openai/openai-python

[12] https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b

[13] We provide detailed settings in Appendix B.3

[14] https://github.com/mjpost/sacrebleu

	Faithfulness	Fluency	Attractiveness
All (= 3)	0.3	0.25	0.17
Majority (≧ 2)	-	-	0.84

Table 3: Inter annotator agreement.

	B-4	R-1	BS	KWD	REG
Unimodal model:					
BM25	5.4	16.1	70.1	**97.0**	45.0
BART	**14.4**	21.4	73.4	75.8	81.0
T5	13.6	**23.0**	**73.8**	89.8	78.5
GPT-3.5	3.5	14.2	64.2	73.9	84.5
GPT-4	4.4	16.4	65.1	78.6	**87.0**
Llama2	4.6	13.6	55.4	72.2	60.0
Multimodal models:					
T5 + {o}	**16.0**	**24.7**	**74.9**	**85.7**	70.0
T5 + {o,l}	15.6	23.3	74.1	84.4	67.5
T5 + {o,l,v}	13.2	23.5	74.1	84.5	**74.0**

Table 4: Results: a **bold** value indicates the best result in each column.

performed manual evaluations for a total of 900 ad texts.

Table 3 shows the inter-annotator agreement (IAA)[15]. As expected, the IAA for attractiveness is the lowest, but when loosened to more than a majority, it is outstandingly high (0.84). This suggests that, while achieving unanimous favorability is challenging, there is a considerable level of consensus on attractiveness.

5.3 Result

The answers corresponding to the RQs listed in §5 are provided below:

A1: Finetuning and few-shot are good performers in intrinsic and extrinsic evaluations, respectively In automatic evaluation, we observe that few-shot learning falls behind finetuning (Table 4). A similar trend can also be observed in the manual evaluation, except for attractiveness (Figure 4 and Figure 5). These series of results highlight the high potential of LLM few-shot for improving quality in *extrinsic* evaluation such as attractiveness and human preference, while finetuning can play an important role in maximizing quality in *intrinsic* evaluation such as automatic scores, faithfulness, and fluency.

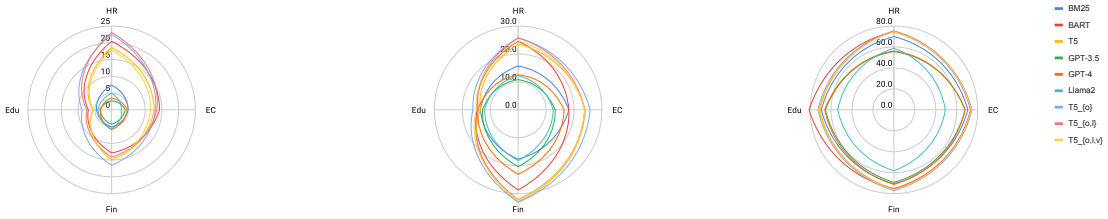
A2: Multimodal information contributes to the quality of generated ad text We observe that incorporating additional features such as OCR-processed text (+ {o}), the LP layout information (+ {o,l}), and LP image features (+ {o,l,v}) improved the quality of generated sentences in terms of faithfulness (4a) and fluency (4b). On the other hand, the incorporation of layout information and visual features into the models does not necessarily improve performance, so methods for model integration require further exploration. The performance drop may be due to image information acting as noise when using the LP Full View directly in this experiment. Therefore, the development of a multimodal system that adaptively accesses only important information from LPs will be a straightforward future work.

A3: Model performance and model rankings vary by industry domain Figure 3 shows the industry-wise evaluation results in each metric. We observe the model performance and rankings vary by industry, especially in B-4 and R-1. In contrast, BS exhibits a relatively stable model ranking across industries. This stability could be attributed, to the embedding-based nature of BS, offering a more flexible interpretation of semantic proximity compared to the surface-based metrics. For a thorough examination of the reliability of each metric in the ATG task, refer to §6.

A4: Some baselines have already reached human-level performers In **faithfulness**, the outputs of the baseline models, with the exception of GPT-3.5 and GPT-4, are more faithful to the input than the human reference (Figure 4a). Note, however, that low faithfulness in human reference does not necessarily mean low quality, since it is known that ad creators use expressions based on their external knowledge to the extent that they can ensure factual consistency with the input in order to enhance fluency and appeal. Non-factual, fake ads can be fatal to advertisers in terms of legal compliance and corporate branding, but it is difficult for a model to perfectly capture real-time product-specific information, such as discount prices and campaign periods. Therefore, one important direction is the development of models with guaranteed faithfulness as a step toward achieving an ATG system with guaranteed factual consistency.

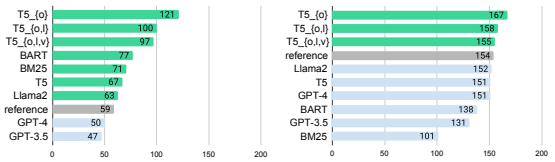
In **fluency**, we can confirm that the human reference has high fluency as a trade-off for low faithfulness, while GPT-4, T5, and Llama2 are almost at the same level as the human reference (Figure 4b). It should also be noted that integrating multimodal

[15]It is based on majority vote and counted as a Tie if they are all split for attractiveness



(a) B-4 (b) R-1 (c) BS

Figure 3: Industry-wise evaluation for each metrics.



(a) Faithfulness (b) Fluency

Figure 4: Human ranking in terms of faithfulness and fluency, respectively.

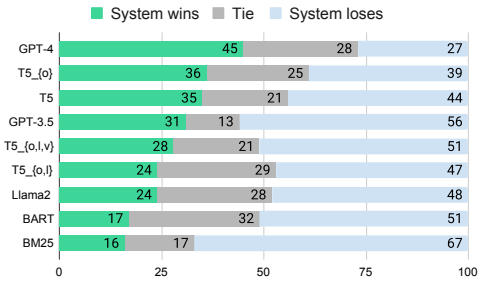


Figure 5: Human preference evaluation for each system output, comparing to a human-created reference.

	Faithfulness		Fluency		Attractivenss	
Metrics	r	ρ	r	ρ	r	ρ
B-4	0.88	0.83	0.53	0.30	-0.12	-0.68
R-1	0.83	0.75	**0.70**	**0.55**	**0.35**	**0.03**
BS	**0.90**	**0.85**	0.67	0.50	0.20	-0.20
GPT-4	0.20	-0.48	-0.22	0.10	-0.47	-1.20

Table 5: System-level meta-evaluation results with Pearson (r) and Spearman (ρ)

information from LP images into the model contributes to generating more fluent ad text.

In **attractiveness**, GPT-4 is already able to generate more attractive ad text for humans than reference (Figure 5). If equivalent (Tie) cases are included, T5 and T5+ {o} also reach the same level as humans. GPT-4 also achieves a sentence-length regulation compliance rate (REG in Table 4), making it a model with high real-world applicability.

6 Meta-evaluation

We investigate the following two questions: (1) *how reliable are the existing metrics for each evaluation aspect?*, and also (2) *can a strong LLM (e.g., GPT-4) be used to be an alternative to human evaluation?* To answer these questions, we performed a meta-evaluation by adding a GPT-4-based evaluator to the set of the metrics used in the experiment in §5. The GPT-4-based evaluator was constructed by giving the same instructions as those given to the human raters in the manual evaluation §5.2.[16].

Table 5 shows that the system-level meta-evaluation results with Pearson (r) and Spearman (ρ). BS and R-1 correlate best with humans for faithfulness and fluency, respectively. On the other hand, it was difficult to replicate the human ranking for attractiveness. This suggests that existing metrics work as intrinsic evaluations, but it is still difficult to use them as a substitute for extrinsic evaluations. The GPT-4 based evaluator had the lowest correlation in any evaluation aspect. This result is inconsistent with Chiang and Lee (2023)'s report that LLM evaluations produce results similar to those of expert human evaluations. One reason for this may be due to domain mismatch, as most of the datasets in the GPT-4 pre-training are general or non-advertising domains (OpenAI, 2023).

7 Conclusion

In this study, we standardized ATG as a cross-application task and developed the first benchmark dataset. Through evaluation experiments using this benchmark, we demonstrated the current status and remaining challenges. ATG is a promising application of NLP and a critical and complex research area for advancing user-centric language technology. We anticipate that the research infrastructure established in this study will drive the progress and development of ATG technology.

[16]The prompts used are presented in Appendix C

Limitations

One of the limitations of this study is that the dataset is only available in Japanese. In particular, the community should also enjoy benchmark datasets in English that are more accessible to researchers and developers around the world. We hope that advertising-related companies who share our vision of building on common datasets to build on the technologies in the field of ATG will follow this research and provide public datasets to the community in the future.

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*, EC '08, page 1–9. Association for Computing Machinery.

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

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the 2005 Document Understanding Conference*.

Ogi Djuraskovic. 2022. Google search statistics and facts 2023 (you must know). Technical report, First Site Guide.

Siyu Duan, Wei Li, Jing Cai, Yancheng He, and Yunfang Wu. 2021. Query-variant advertisement text generation with association knowledge. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 412–421. Association for Computing Machinery.

Atsushi Fujita, Katsuhiro Ikushima, Satoshi Sato, Ryo Kamite, Ko Ishiyama, and Osamu Tamachi. 2010. Automatic generation of listing ads by reusing promotional texts. In *Proceedings of the 12th International Conference on Electronic Commerce: Roadmap for the Future of Electronic Business*, ICEC '10, page 179–188. Association for Computing Machinery.

Hiroyuki Fukuda. 2019. Keyword conditional variational autoencoder for advertising headline generation (in japanese). In *The 33rd Annual Conference of the Japanese Society for Artificial Intelligence,*, pages 2L4J903–2L4J903.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

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*, KDD '19, page 2269–2277. Association for Computing Machinery.

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. Association for Computational Linguistics.

Yashal Shakti Kanungo, Gyanendra Das, Pooja A, and Sumit Negi. 2022. Cobart: Controlled, optimized, bidirectional and auto-regressive transformer for ad headline generation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3127–3136. Association for Computing Machinery.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Haonan Li, Yameng Huang, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin, and Nan Duan. 2022. CULG: Commercial universal language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 112–120. Association for Computational Linguistics.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. Association for Computational Linguistics.

Mary Meeker and Liang Wu. 2018. Internet trends 2018. Technical report.

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.

Soichiro Murakami, Sho Hoshino, and Peinan Zhang. 2023. Natural language generation for advertising: A survey.

Soichiro Murakami, Sho Hoshino, Peinan Zhang, Hidetaka Kamigaito, and Manabu Takamura, Hiroyaand Okumura. 2022a. Lp-to-text: Multimodal ad text generation (in japanese). In *The 28th Annual Conference of the Association for Natural Language Processing*.

Soichiro Murakami, Peinan Zhang, Sho Hoshino, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2022b. 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. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu 0003, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1).

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Walter Dill Scott. 1903. *The theory of advertising*. Small, Maynard and Company.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Stamatina Thomaidou, Ismini Lourentzou, Panagiotis Katsivelis-Perakis, and Michalis Vazirgiannis. 2013. Automated snippet generation for online advertising. CIKM '13, page 1841–1844. Association for Computing Machinery.

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. 2023. Llama 2: Open foundation and fine-tuned chat models.

Taifeng Wang, Jiang Bian, Shusen Liu, Yuyu Zhang, and Tie-Yan Liu. 2013. Psychological advertising: Exploring user psychology for click prediction in sponsored search. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 563–571. Association for Computing Machinery.

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 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*.

Penghui Wei, Xuanhua Yang, ShaoGuo Liu, Liang Wang, and Bo Zheng. 2022. CREATER: CTR-driven advertising text generation with controlled pre-training and contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry*

Track, pages 9–17. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263. Association for Computational Linguistics.

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*, WWW '20, page 1366–1377. Association for Computing Machinery.

Chao Zhang, Jingbo Zhou, Xiaoling Zang, Qing Xu, Liang Yin, Xiang He, Lin Liu, Haoyi Xiong, and Dejing Dou. 2021. Chase: Commonsense-enriched advertising on search engine with explicit knowledge. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4352–4361. Association for Computing Machinery.

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 Annotation Guideline

The main instructions given to the annotators were as follows:

1. Consider the search keyword as the user's intent.
2. Create an advertisement that is consistent with the product/service description in the LP.
3. Ensure that the length of the advertisement is within 15 full-width characters [17].
4. These instructions were provided to guide the annotators in creating the additional reference advertisements.

B Details on experimental setup for each baseline models

B.1 BM25

We used the BM25 to rank sentences of the source document given a query and took the most relevant sentence as the generated ad text. For implementation, we used the `rank_bm25` toolkit [18].

B.2 T5 and BART

We fine-tuned each pre-trained model on the training dataset to create our baseline models. Specifically, we used a pre-trained model `japanese_bart_base_2.0` from Kyoto University's Japanese version of BART [19] as the basis for our BART-based baseline model. For the T5-based baseline model, we used a pre-trained model `sonoisa/t5-base-japanese` [20]. The specific hyperparameters and other experimental details are reflected in Table 6.

B.3 Multimodal models

Figure 6 presents an overview of incorporating the LP information into the T5-based model. [21]. As an input, we used three sets of token sequences, the LP descriptions x^{des}, user queries x^{qry}, and each OCR token sequence x_i^{ocr} of the rectangular region set $R = \{r_i\}_{i=1}^{|R|}$ obtained by OCR from the LPs, where each token sequence x^* is $x^* = (x_i^*)_{t=i}^{|R|}$. Furthermore, the layout $C = c_i{}_{i=1}^{|R|}$ and image information $I = I_i{}_{i=1}^{|R|}$ for the rectangular region set R was used. Here, c_i denotes $(x_i^{\mathsf{min}}, x_i^{\mathsf{max}}, y_i^{\mathsf{min}}, y_i^{\mathsf{max}}) \in \mathbb{R}^4$ as shown in Figure 6.

Next, we explicitly describe each embedding (Figure 6) as follows:

Token embedding Each token sequence x^* was transformed into an embedding sequence t^* before being fed into the encoder. Here, D denotes the embedding dimension.

Segment embedding The encoder distinguishes the region of each token sequence x^*. For example, for a token sequence x^{des}, we introduced $s^{des} \in \mathbb{R}^D$.

Visual embedding We introduced an image I_i for each rectangular region r_i to incorporate visual information from the LP, such as text color and font. More specifically, the obtained image I_i was

[17] This follows the guidelines for headline text in Google Responsive Search Ads (https://support.google.com/google-ads/answer/12437745).

[18] https://github.com/dorianbrown/rank_bm25

[19] https://github.com/utanaka2000/fairseq/tree/japanese_bart_pretrained_model

[20] https://huggingface.co/sonoisa/t5-base-japanese

[21] Note that the model constructed for this experiment, shown in Figure 6, is not the proposed model, but a baseline model created according to Murakami et al. (2022a)

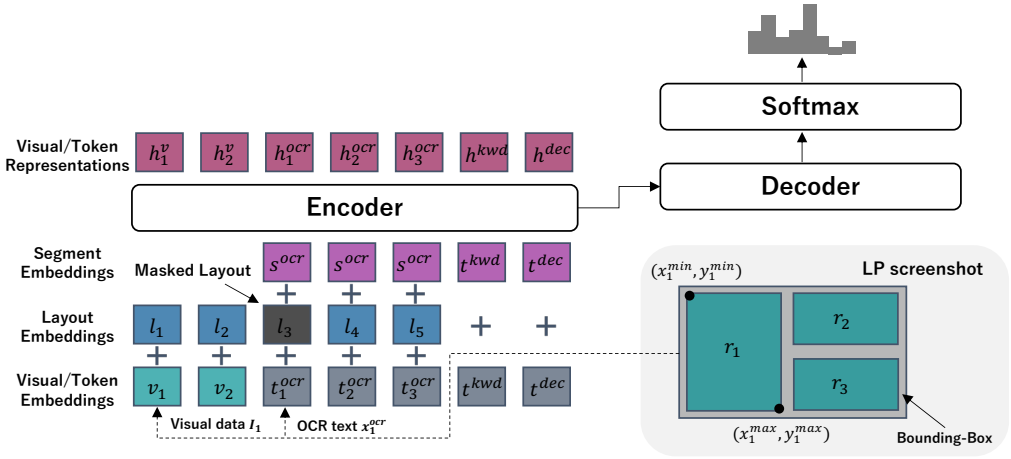


Figure 6: An overview of the model incorporating LP information, following Murakami et al. (2022a).

resized to 128 × 32 (width × height). The CNN-based feature extraction was employed to create visual features $v_i \in \mathbb{R}^D$.

Layout embedding In the LP, the position and size of the letters played crucial roles. We input the layout c_i of a rectangular region r_i into the MLP to obtain $l_i \in \mathbb{R}^D$.

Using the above embeddings, we generated the encoder inputs, as shown in Figure 6. This study investigated the contribution of each type of multimodal information to the overall performance. We incorporated the following three types of multimodal information into the model architecture in Figure 6: LP OCR text (`lp_ocr;o`), LP layout information (`lp_layout;l`), and LP BBox image features (`lp_visual;v`).

Hyperparameters We present the hyperparameters used during the training of both models in Table 6. For the maximum sequence length in T5, it was set to 712 only for the model using LP bounding box image features (`+{o,l,v}`), while all other models were set to 512. Furthermore, early stopping was applied if the loss on the development set deteriorated for 3 consecutive epochs in the case of T5, and 5 consecutive epochs in the case of BART.

B.4 GPT-3.5, GPT-4, and Llama2

For GPT-3.5, GPT-4, and Llama2, the baseline models were constructed by 3-shot in-context learning, respectively. The prompts used to build these models are provided in Table 7.

C Prompts for GPT-4 evaluator

The GPT-4-based evaluator was constructed by giving the same instructions as those given to the human raters in the manual evaluation §5.2. We present the prompts we used for faithfulness, fluency, and attractiveness in Tables 8, Table 9, and Table 10, respectively.

Hyperparameters	Values （BART / T5）
Models	`japanese_bart_base_2.0` / `t5-base-japanese`
Optimizer	Adam (Kingma and Ba, 2015)
Learning rate	3e-4
Max epochs	20
Max length	512 / 712 （T5+`{o,l,v}` only ）

Table 6: Hyperparameters.

Based on the given search query and text, please create an advertisement that appeals to users in 15 words or less.

Search Query: bridal fair Yokohama
Document: Official website of "The House Yokohama Marine Tower Wedding", a wedding venue at Yokohama Marine Tower adjacent to Yamashita Park. One couple can rent out the Yokohama Marine Tower, which overlooks Minato Mirai, and have a wedding ceremony that is unique to them.
Output: Yokohama wedding THE HOUSE open

Search Query: window cleaning
Documents: Compare window and sash cleaning prices, quotes, and reviews at Kurashi no Market. Easily book reputable window and sash cleaning professionals online! [Guaranteed!]
Output: [Official] Kurashino Market

Search Query: jobs osaka 50s
Documents: Find the right job for you at Recruit's job search and job information site! Rikunabi NEXT is a job search and recruitment information site that supports your job search with useful contents such as job scout function and know-how on job change.
Output: Many senior jobs are available

Search query: $\{query\}$
Documentation: $\{description\}$
Output:

Table 7: Prompts used for ATG model based on LLMs (GPT-3.5, GPT-4, and Llama2), translated into English for visibility.

Please answer "1" if the question text implies the ad text and "0" if it does not.

Question text: [A calm daily life begins with a regular diet] Self-care for common female problems/regular delivery costs about 81 yen a day.
Ad text: Peaceful everyday life
Answer: 1

Question text: [A calm daily life begins with a regular diet] Self-care for common female problems/regular delivery costs about 81 yen a day.
Ad text: [Official] Daily diet
Answer: 0

Question: How to recover/restore data from an external hdd?
Ad text: 0 yen for the initial cost
Answer: 0

Question: $\{description\}$
Ad text: $\{adtext\}$
Answer:

Table 8: Prompt used for GPT-4 evaluator for faithfulness, translated into English for visibility.

Please answer "1" for the following ad text if the content is understandable and natural, and "0" otherwise.

Ad text: You get muji miles every year.
Answer: 1

Text: [Official] marriveil
Answer: 1

Ad text: ujipassport app
Answer: 0

Ad text: $\{adtext\}$
Answer:

Table 9: Prompt used for GPT-4 evaluator for fluency, translated into English for visibility.

Assuming a Google search for the following keywords, please compare ad text A and ad text B and answer "A" or "B" for the one you are more interested in. If the attractiveness is the same, please answer "C".

Keyword: employment information
Ad text A: [Official] TOYOTA / Recruitment of periodic employees
Ad text B: [Official] TOYOTA / Periodic Employee Recruitment
Answer: A

Keyword: recommended medical insurance
Ad text A: Nippon Life Group Medical Insurance
Ad text B: Online Medical Insurance
Answer: B

Keyword: cancer hospital visit insurance
Ad text A: Sony Assurance's medical insurance
Ad text B: Aflac medical insurance
Answer: C

Keyword: $\{query\}$
Ad text A: $\{reference\}$
Ad text B: $\{system\}$
Answer:

Table 10: Prompt used for GPT-4 evaluator for attractiveness, translated into English for visibility. The examples of prompts were selected by sampling from cases in which the evaluators' opinions were in total agreement during the manual evaluation.