AdTEx: Evaluating Large Language Models for Diverse Advertisement Template Expansion

Hiroto Otake[†], **Peinan Zhang**[‡], **Yusuke Sakai**[†], **Hiroki Ouchi**[†‡], **Taro Watanabe**[†]
[†]Nara Institute of Science and Technology (NAIST) [‡]CyberAgent AI Lab
otake.hiroto.od2@naist.ac.jp, zhang_peinan@cyberagent.co.jp
{sakai.yusuke.sr9, hiroki.ouchi, taro}@is.naist.jp

Abstract

Sponsored Search Ads, which are displayed alongside search results, must satisfy strict constraints on their wording and content, making manual creation both time-consuming and labor-intensive. Although template-based approaches enable efficient ad creation, maintaining a diverse set of advertisements is equally important, as repeatedly displaying similar ads can create a negative impression among users. In this paper, we propose AdTEx, a benchmark framework for evaluating the ability of large language models (LLMs) to expand existing advertisement templates while promoting diversity. AdTEx is built on a dataset curated from real-world template-based advertising campaigns in Japan, consisting of 151,103 headlines and 12,076 templates, and evaluates the ability of LLMs to generate diverse template expansions. Experimental results show that, although further improvements are needed in terms of content consistency, semantic similarity, and compliance with advertising format requirements, some LLMs can generate expanded templates that satisfy character-length constraints while increasing surface diversity.

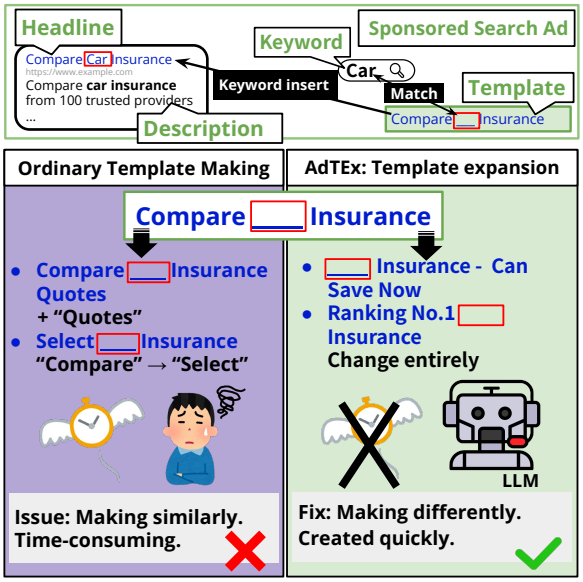


Figure 1: Overview of **AdTEx**. Sponsored Search Ads display headlines and descriptions alongside search results. Given a headline group and an existing advertisement template as a format exemplar, AdTEx evaluates whether LLMs can generate diverse template expansions while preserving the original advertising intent and satisfying practical ad delivery constraints, such as character-length and formatting requirements.

1 Introduction

Sponsored Search Ads (ads), which are displayed alongside search results, are widely used in web advertising to promote products and services to users based on their search queries. Such ads must satisfy strict requirements on their wording and content while remaining relevant to the advertised product and attractive to potential customers (Murakami et al., 2023). Moreover, manual ad creation is both time-consuming and labor-intensive because advertisers must produce numerous ad variations under practical constraints, including character-length limits, formatting requirements, and advertising compliance requirements. Therefore, Automatic Ad Text Generation (ATG) has attracted considerable attention as a means of reducing this workload.

In the ATG domain, thanks to recent advances in natural language generation (NLG) with Large Language Models (LLMs), the fluency and quality of automatically generated advertisements have substantially improved (Kamigaito et al., 2021; Wang et al., 2021; Jin et al., 2023; Kamigaito et al., 2024; Wang et al., 2025; Jiang et al., 2025; Chen et al., 2025; Zhang et al., 2025). Nevertheless, generated advertisements still require manual verification before deployment to ensure compliance with advertising requirements. This verification process becomes a major bottleneck when advertisers must quickly create large numbers of ads, for example, for product catalogs containing many low-priced items, such as screws and small accessories, or for

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 1249–1262
October 17-21, 2026 ©2026 Association for Computational Linguistics

location-specific campaigns, where each product or location has a limited advertising budget but still requires a separate ad.

To reduce this cost, template-based advertisement delivery has become widely adopted (Kang et al., 2025; Lee et al., 2025; Zhuang and Sun, 2025). In this approach, advertisement headlines and phrases are manually designed as templates, while product-specific information is represented as placeholders. By simply filling in these placeholders with product names or other attributes, advertisers can rapidly generate large numbers of advertisements that satisfy character-length, formatting, and compliance requirements at low cost. However, creating high-quality templates remains a manual process that requires substantial effort.

Furthermore, relying on a limited set of templates reduces the diversity of delivered advertisements. Repeated exposure to similar advertisements can negatively affect users' impressions and reduce advertising effectiveness (Pechmann and Stewart, 1988; Abrams and Vee, 2007). Therefore, generating or expanding existing templates while preserving their advertising intent has become an important challenge. LLMs offer a promising solution for template expansion or generation. Therefore, it is essential to systematically evaluate their ability to generate diverse yet practical advertisement templates. However, systematic benchmark datasets and standardized evaluation frameworks remain scarce because advertising templates and delivery systems are proprietary.

To address this challenge, we propose AdTEx, a benchmark framework for systematically evaluating the ability of LLMs to expand advertisement templates while promoting diversity, as illustrated in Figure **??**. AdTEx is constructed from Japanese real-world template-based advertising campaigns and consists of 151,103 headlines and 12,076 templates. The benchmark evaluates whether expanded templates increase surface diversity while preserving semantic consistency and satisfying practical advertising constraints, including character-length and formatting requirements. In addition, we provide a baseline for expanding diverse advertisement templates using LLMs. Our experiments show that some LLMs successfully generate template expansions with greater surface diversity while satisfying character-length constraints. However, further improvements are required in content consistency, semantic similarity, and compliance with advertisement formatting requirements, highlighting the challenges that remain for practical deployment.

To summarize, our contributions are as follows:

- We define the task of diverse advertisement template expansion and propose **AdTEx**, a benchmark framework for systematically evaluating the template expansion capabilities of LLMs.
- We construct a benchmark dataset from real-world template-based advertising campaigns in Japan, consisting of 151,103 headlines and 12,076 templates, providing a valuable resource for future research on advertisement template generation and expansion.
- We establish a prompt-based baseline for diverse advertisement template expansion and benchmark multiple LLMs. Our experimental results identify the remaining challenges for practical deployment and provide insights for future research and model development.

2 Background and Related Work

Responsive search ads. Responsive Search Ads (RSAs)[1] display headlines and descriptions alongside search results according to a user's search query. Rather than relying on a single advertisement, RSAs automatically select combinations of candidate headlines and descriptions to maximize relevance and effectiveness. Furthermore, keyword insertion[2] dynamically replaces placeholders with search-query-dependent keywords, enabling scalable advertisement delivery. In this work, we treat headlines containing keyword-insertion placeholders as **advertisement templates**. These advertisement templates are primarily used for products with many variants, such as accessories that differ only in color or model number, or for localized advertisements in which only the location name is dynamically inserted while the remaining content is shared across advertisements. Such template-based advertisements are particularly common in large-scale, low-margin advertising campaigns.

Advertisement generation. Automatic Ad Text Generation (ATG) has been widely studied to reduce the manual effort required for creating advertisements. Early approaches focused on example-based generation and template filling (Bartz et al.,

[1] https://support.google.com/google-ads/answer/7684791?hl=en [Accessed: 15th July, 2026]

[2] https://support.google.com/google-ads/answer/2454041?hl=en [Accessed: 15th July, 2026]

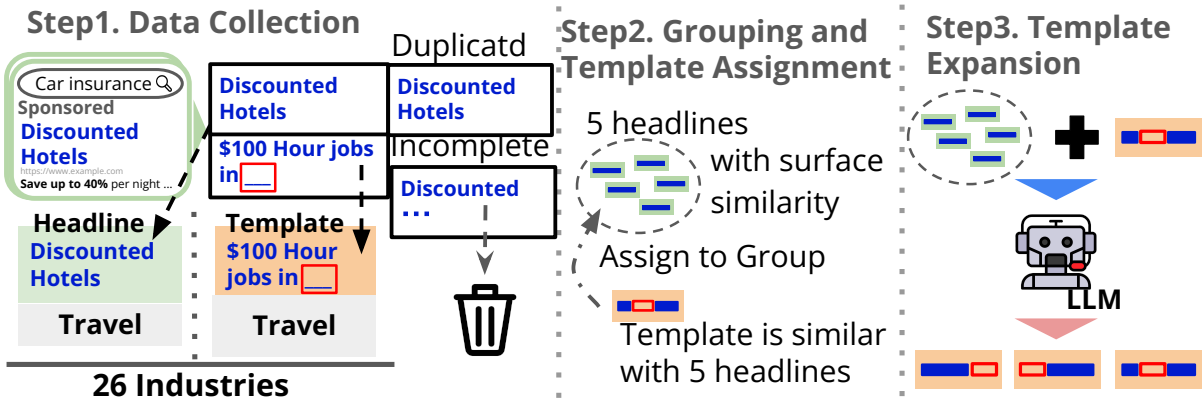


Figure 2: The template expansion framework consists of 3 steps. Step 1: We collected and categorized headlines by industry. Step 2: We clustered headlines within each industry and assigned templates to each headline group. Step 3: We expanded templates by extracting common expressions from multiple headlines using existing templates. multiple headlines using existing templates. multiple headlines using existing templates.

2008; Thomaidou et al., 2013b; Alnajjar and Toivonen, 2021), information extraction from product documents (Thomaidou et al., 2013a; Golobokov et al., 2022), and summarization or rewriting (Fujita et al., 2010; Hughes et al., 2019; Mishra et al., 2020). More recently, LLMs have substantially improved the fluency and quality of automatically generated advertisements (Kamigaito et al., 2021, 2024; Wang et al., 2025; Jiang et al., 2025; Chen et al., 2025). Existing studies have primarily focused on generating entire advertisements or improving generation quality, whereas our work focuses on evaluating the ability of LLMs to expand existing advertisement templates while promoting diversity under practical advertising constraints.

Advertisement generation benchmarks. CAMERA (Mita et al., 2024), CAMERA^3 (Inoue et al., 2024), FaithCAMERA (Kato et al., 2024), and LCTG Bench (Kurihara et al., 2025) evaluate the ability of LLMs to generate advertisement text under practical advertising scenarios, while AdTEC (Zhang et al., 2025) provides a unified benchmark for assessing advertisement quality from multiple operational perspectives. Notably, these benchmarks are developed for Japanese advertisements, reflecting the active research community and the availability of high-quality advertising resources in this domain. In contrast, our work focuses on advertisement template expansion, which aims to generate diverse variants of existing human-written advertisement templates, thereby improving the efficiency of real-world advertising workflows by reducing the manual effort required to create and maintain advertisement templates.

3 AdTEx

We propose AdTEx, a benchmark framework for systematically evaluating the ability of LLMs to perform diverse advertisement template expansion under practical advertising constraints. As illustrated in Figure 2, AdTEx is constructed from real-world advertisement delivery logs. We first collect advertisement headlines and organize them by industry. Within each industry, semantically related advertisements are grouped together, and a representative advertisement template is assigned to each group. Given the resulting benchmark, LLMs are prompted to generate diverse template expansions by leveraging existing templates and common expressions extracted from multiple advertisements. The generated templates are then evaluated in terms of diversity, semantic consistency, and compliance with practical advertising constraints.

3.1 Dataset Construction

To construct AdTEx, we collected advertisement data that had been used in real-world Japanese Sponsored Search Ad campaigns. Because advertisement uniqueness is important for attracting users' attention (Mafael et al., 2021), duplicated headlines were removed from the collected data. We also excluded incomplete advertisements whose headlines ended with an ellipsis, indicating truncated content. The resulting advertisements contain both reusable templates and instantiated headlines. We therefore separated them into **templates** and **headlines** according to their structural characteristics. To facilitate industry-wise evaluation, each headline was further assigned to one

Industry	Headline	Template
Academic interests	5,805	1,028
Arts and culture	317	12
Automotives	9,266	527
Books and literature	22,884	687
Business and finance	5,874	284
Careers	8,010	1,436
Family and relationships	3,096	38
Food and drinks	2,166	49
Health	522	320
Healthy living	8,720	750
Hobbies and interests	134	58
Home and garden	1,445	214
Movies	1,354	15
Music and audio	555	22
News and politics	2,188	15
Personal finance	10,978	2,344
Pets	3,141	88
Pharmaceuticals, conditions, and symptoms	248	17
Real estate	10,682	509
Shopping	19,126	482
Sports	1,210	27
Style and fashion	7,750	198
Technology and computing	1,035	211
Television	810	9
Travel	12,965	2,632
Video gaming	10,822	104
Total	**151,103**	**12,076**

Table 1: Distribution of advertisement headlines and templates across the 26 IAB industry categories.

of the 26 IAB categories.[3] Industry classification was performed using a Japanese BERT model (*tohoku-nlp/bert-base-japanese-v3*) fine-tuned on the *shishir-dwi/News-Article-Categorization_IAB* dataset, which contains 871,909 samples. The dataset was split into training, validation, and test sets with a ratio of 8:1:1. The model was trained with a batch size of 128, a learning rate of 2×10^{-5}, and a weight decay of 0.01 following the Hugging Face Trainer (Wolf et al., 2020). Early stopping with a patience of three epochs was applied. Table 1 shows the resulting category distribution. After preprocessing, AdTEx consists of 151,103 headlines and 12,076 advertisement templates.

3.2 Headline Grouping

To construct template expansion instances, we grouped semantically similar advertisement headlines and assigned an existing advertisement template to each group as an exemplar of the desired output format. Specifically, we grouped every five headlines by maximizing their mutual similarity measured with Pairwise-BLEU (Shen et al., 2019). We then assigned the most appropriate template by computing the average 4-gram BLEU (Papineni et al., 2002) between each headline group and every candidate template, selecting the template with the highest average BLEU score. Both Pairwise-BLEU between headlines and BLEU between headlines and templates were computed using 4-gram BLEU. Following this procedure, we obtained 132,863 headline groups with their corresponding advertisement templates.

[3] https://www.iab.com/

Prompt for template expansion

Based on multiple specific ad texts, create several ad text templates by extracting their common structures.

The templates must strictly adhere to the following conditions:

- Templates should represent only the structure of the ad text and must not include any specific information such as actual company names, service names, or place names.
- Fixed parts that cannot be replaced must be within a total of 10 full-width characters.
- Replaceable sections must use {}, and include 1 example word within the {}.
- Do not include any explanatory text or supplementary information; return only the template body.
- When multiple ad examples are provided, consolidate them into a single template with the most versatile common structure.

[Output Format]
Refer to the expected output format and return only the template text.

[Headline examples]

- For Parents Raising Children [Full-Time Job Openings]
- Co-op Mutual Aid] For Pregnant Women
- [For Those Job Hunting]
- For Those Considering Floor Waxing
- Job Openings Suitable for Working Parents

[Expected output format]
For {keyword:Those Seeking Technical} Support

Figure 3: Prompt used for advertisement template expansion. English translation is shown here for readability. Appendix A provides the original Japanese prompt.

3.3 Template Expansion

Following CAMERA (Mita et al., 2024) and CAMERA[3] (Inoue et al., 2024), we construct a prompt that provides the character-length constraints and an existing advertisement template as an example of the desired output format. As illustrated in Figure 3, the prompt further includes five

	Before	After	Yields	Orig.×
Original Template	12,076		100%	1.00×
GPT-4o-mini	71,464	60,983	85.33%	6.05×
Gemini-2.0-Flash	95,396	85,521	89.65%	8.08×
CALM2-7B	157,359	14,516	9.22%	2.20×
CALM3-22B	77,549	55,094	71.04%	5.56×
Llama-3-ELYZA-JP-8B	107,504	19,361	18.01%	2.60×
Llama-2-ELYZA-JP-7B	22,676	5,934	26.17%	1.49×
Llama-2-ELYZA-JP-13B	12,751	5,416	42.48%	1.45×

Table 2: Statistics of generated templates before and after filtering. Yield denotes the percentage retained after filtering, and Orig.× denotes the expansion ratio with respect to the number of original templates.

delivered advertisement headlines and instructs the model to extract their shared structure and generate diverse advertisement templates. The five headlines are randomly sampled from the corresponding headline group constructed in Section 3.2.

During generation, the model produces multiple templates in a single response. The generated templates are separated using newline characters and treated as individual candidates. Because a valid advertisement template must contain at least one replacement slot that can be instantiated according to a search query during ad delivery, candidates without any replacement slots are discarded.

4 Preliminary Experiment

Settings. We evaluated a total of seven LLMs using AdTEx: Llama-3-ELYZA-JP-8B[4], Llama-2-ELYZA-JP-7B[5], -13B[6], CALM3-22B[7], CALM2-7B[8], Gemini-2.0-Flash[9], and GPT-4o-mini[10]. For open-weight models, we use an NVIDIA A6000 GPU. Following Sakai et al. (2024a), we set the temperature to 0 for Gemini-2.0-Flash and GPT-4o-mini, while using greedy decoding with a fixed random seed for the open-weight models.

Results. Table 2 summarizes the total number of templates generated by each model after filtering outputs that do not contain query-dependent replacement slots, together with the number of original advertisement templates. Table 3 further reports the corresponding statistics for each industry category. The results indicate that current LLMs are capable of expanding advertisement templates to a certain extent. However, substantial differences are observed across models. Gemini-2.0-Flash expands the original template inventory by approximately $8\times$, whereas Llama-2-ELYZA-JP-13B achieves only $1.45\times$ expansion. In contrast, although CALM2-7B generates the largest number of candidate templates, most are discarded during post-processing because they do not satisfy the template requirements, resulting in a yield of only about 9%. These findings suggest that successful advertisement template expansion depends not only on generation capability but also on the ability to accurately follow structural instructions and generate valid advertisement templates.

[4] elyza/Llama-3-ELYZA-JP-8B
[5] elyza/ELYZA-japanese-Llama-2-7b-instruct
[6] elyza/ELYZA-japanese-Llama-2-13b-instruct
[7] cyberagent/calm3-22b-chat
[8] cyberagent/calm2-7b-chat
[9] Gemini/gemini-2.0-flash
[10] OpenAI/gpt-4o-mini-2024-07-18

5 Evaluation Metrics

Next, we evaluate the quality of the advertisement templates generated in Section 4.

5.1 Automatic Evaluation Metrics

We evaluate the framework in terms of template expansion, surface diversity among generated templates, and suitability for ad delivery. The augmentation rate (**AUG**) measures how much the number of templates increases compared to the original data. For diversity evaluation, we employed Pairwise-BLEU (**pBLEU**) and Pairwise-BLEURT (**pBLEURT**) (Sakai et al., 2025), calculated by the `mbrs` library (Deguchi et al., 2024). We interpret pBLEU as a measure of surface similarity and pBLEURT as a measure of semantic similarity. Accordingly, lower pBLEU indicates greater surface diversity, whereas higher pBLEURT indicates greater semantic similarity among generated outputs. We computed weighted averages over industries, using the number of data points per industry for pBLEU and pBLEURT. Industries with zero templates after filtering out those without replacement slots were excluded from the diversity evaluation.

Furthermore, to evaluate whether the generated templates are suitable for ad delivery, we introduce the compliance rate with the character length limit (**REG**) (Mita et al., 2024) and the keyword insertion format constraint (**KREG**). REG indicates the proportion of headlines that satisfy the character length constraints, defined as within 15 full-width or 30 half-width characters. KREG represents the proportion of templates that satisfy the required format for keyword insertion. Specifically,

	GPT-4o	Gemini	CALM2-7B	CALM3-22B	ELYZA-JP-8B	ELYZA-JP-7B	ELYZA-JP-13B	Original
Academic interests	2658	3159	448	2352	738	425	264	1028
Arts and culture	204	235	57	190	42	0	3	12
Automotives	3957	4771	542	3604	1542	404	292	527
Books and literature	5850	10878	5008	4316	2081	27	230	687
Business and finance	3423	3820	385	3252	1188	552	374	284
Careers	4294	5447	808	4265	1682	673	370	1436
Family and relationships	1673	1691	182	1291	510	132	95	38
Food and drinks	940	1154	140	862	329	139	206	49
Health	292	362	48	216	96	68	17	320
Healthy living	5009	6042	795	4069	2136	188	438	750
Hobbies and interests	95	105	22	101	38	5	6	58
Home and garden	737	867	441	577	281	7	80	214
Movies	525	772	131	857	208	38	49	15
Music and audio	361	461	243	373	168	10	60	22
News and politics	1026	1215	70	679	161	95	22	15
Personal finance	4917	6696	739	3393	1507	544	425	2344
Pets	1489	2155	359	1283	476	147	113	88
Pharmaceuticals, ...	152	162	26	152	50	2	6	17
Real estate	3147	5354	370	1643	343	212	100	509
Shopping	6695	11030	1213	6826	1729	1183	640	482
Sports	292	383	40	243	43	4	6	27
Style and fashion	4432	5886	517	5161	1087	41	659	198
Technology and computing	718	850	92	881	184	115	93	211
Television	287	406	94	411	160	0	20	9
Travel	4475	6419	1245	3907	1404	492	556	2632
Video gaming	3335	5201	501	4190	1178	431	292	104

Table 3: Number of generated templates after filtering and original templates for each industry.

the format includes curly braces containing content separated by a colon, where the part before the colon includes a predefined term such as *Keyword*, and the part after the colon provides an example word. Alternatively, the content within the curly braces may follow a pattern such as *Location(City)* or *Location(Area)*, where *Location* is followed by a region-indicating word in parentheses. In addition, the template must contain a non-symbol character outside the curly braces. Furthermore, for the evaluation of pBLEURT and REG, we evaluated headlines made of the expanded templates. As baselines for these metrics, we used the results of the original templates for each metric.

5.2 Manual Evaluation Metrics

To verify the practical usability of the generated advertisement templates, we conducted a manual evaluation. We first generated advertisement headlines from each template using Gemini-2.0-Flash. Figure 4 shows the binary screening prompt for assessing an already instantiated headline, rather than a prompt for generating its text. Among all generated templates, 2,619 templates that produced sufficiently specific headlines were selected for evaluation. Advertising domain experts evaluated the generated headlines and their corresponding templates using three binary criteria. **Grammar** measures whether the generated headline contains

Prompt for Ad Template Extraction

You are an assistant responsible for evaluating advertising text.
You will be given an advertisement generated from a template by the user.

Your task is as follows:

- Determine whether the advertisement is sufficiently specific and semantically valid as an advertisement.
- Return False if the advertisement is semantically invalid, unnatural, or too abstract.
- Return False if the advertisement contains unnecessary or inappropriate terms for advertising.
- Return True only if the advertisement is natural, specific, and suitable for use as an advertisement.

Follow the output format below:

- Respond with either True or False.
- You must return only True or False.

[Headline generated from the template]
{headline}

Figure 4: Prompt used for template filtering. The original Japanese prompt is provided in Appendix A; the English translation is shown here for readability.

	Automatic Evaluation					Manual Evaluation (↑)		
	AUG (↑)	pBLEU (↓)	pBLEURT (↑)	REG (↑)	KREG (↑)	Grammar	Coherence	Rational
Original	100.00	11.26± 5.55	11.13± 6.29	99.99	85.19	100.00	100.00	100.00
GPT-4o-mini	504.99	8.15± 3.39	10.69± 5.30	88.38	75.67	91.15	50.08	**85.33**
Gemini-2.0-Flash	**708.19**	6.22± 2.00	**11.31**± 5.30	**91.82**	17.32	90.18	69.03	83.94
CALM2-7B	120.21	8.46± 3.31	10.60± 5.98	82.07	50.85	**93.76**	**73.24**	84.01
CALM3-22B	456.23	8.60± 3.96	9.02± 4.57	84.03	77.25	91.98	61.35	83.61
Llama-3-ELYZA-JP-8B	160.33	7.10± 3.42	7.86± 4.45	86.97	65.27	77.79	44.93	79.02
Llama-2-ELYZA-JP-7B	49.14	11.92± 5.68	8.88± 5.79	85.34	**87.82**	56.81	31.00	41.95
Llama-2-ELYZA-JP-13B	44.85	**5.90**± 3.25	8.61± 5.00	41.18	59.08	80.50	57.33	71.65

Table 4: Evaluation result. **Bold** is the highest value, underline is the second one. ± indicates the standard deviation.

grammatical errors. **Coherence** measures whether the generated headline is semantically coherent and natural. **Rational** measures whether the advertisement template itself is semantically coherent and suitable for reuse as a template.

Moreover, commercial advertising platforms impose strict grammatical and content requirements on advertisements[11]. Because all collected advertisement templates had been used in real-world advertising campaigns, we consider them to satisfy these requirements and report the original templates as the human evaluation baseline (100%).

6 Results

Table 4 shows the overall results of the evaluation.

Automatic evaluation. Focusing on AUG in Table 4, most models achieve higher than the baseline, suggesting that AdTEx can generate more templates than the original ones. Focusing on pBLEU and pBLEURT in Table 4, while most models also obtain lower pBLEURT scores, indicating weaker semantic consistency than the original templates, most models achieve lower pBLEU scores than the baseline, indicating greater surface diversity. Focusing on REG and KREG in Table 4, most models achieve over 80% for REG. In contrast, KREG shows variation across models. These results suggest that while models tend to adhere to character length limits, their ease of compliance with specific formatting requirements varies.

Manual evaluation. The manual evaluation results in Table 4 indicate that most models achieve around 80% or higher in Grammar and Rational, approaching the baseline. While Coherence reaches at most approximately 73%, suggesting that content coherence requires improvement compared to grammatical correctness and template validity.

[11] https://support.google.com/adspolicy/answer/6021546 [Accessed: 15th July, 2026]

	pBLEU		pBLEURT	
	r	p-value	r	p-value
GPT-4o-mini	0.18	< 0.001	0.25	< 0.001
Gemini-2.0-Flash	0.25	< 0.001	0.42	< 0.001
CALM2-7B	0.08	< 0.001	0.16	< 0.001
CALM3-22B	0.20	< 0.001	0.28	< 0.001
Llama-3-ELYZA-JP-8B	0.20	< 0.001	0.25	< 0.001
Llama-2-ELYZA-JP-7B	0.02	0.20	0.09	< 0.001
Llama-2-ELYZA-JP-13B	0.06	< 0.001	0.08	< 0.001

Table 5: Correlation results between the pBLEU of headlines and the pBLEU and pBLEURT of the generated templates. r denotes Pearson's correlation coefficient.

	No Word	Insert	Both	Total
GPT-4o-mini	11993	2672	173	14838
Gemini-2.0-Flash	1478	62714	6518	70710
CALM2-7B	7051	31	53	7135
CALM3-22B	12325	195	13	12533
Llama-3-ELYZA-JP-8B	5162	1114	448	6724
Llama-2-ELYZA-JP-7B	710	8	5	723
Llama-2-ELYZA-JP-13B	2077	106	33	2216

Table 6: Number of instances for each error type by model. Total represents the total number of errors.

7 Analyses and Discussions

Association between headline surface similarity and generated-template similarity. Since higher surface similarity reflects more similar vocabulary and greater semantic similarity due to shared product details (Aoki et al., 2025), we compute the correlation between the Pairwise-BLEU of the headlines and the pBLEU and pBLEURT of the generated templates to analyze their relationships. Although most models show no significant correlation between Pairwise-BLEU and pBLEU, weak to moderate correlations are observed, suggesting that surface-similar headlines lead to the generation of semantically similar templates.

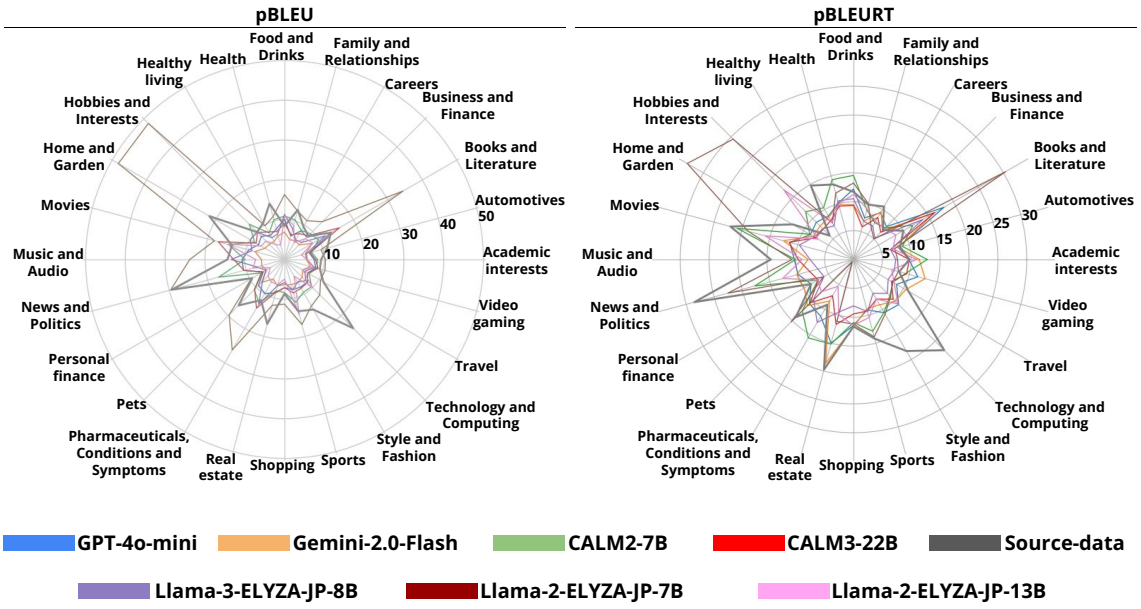


Figure 5: Industry-wise diversity evaluation results.

Error analysis for template format. We analyzed the templates generated by each model that do not conform to the required format. Table 6 presents the number of instances of each error type by model. We denote cases where no non-symbol characters appear outside the curly braces as **No Word**, cases where unsupported insertion phrases are generated as **Invalid Insert**, and cases containing both failures as **Both**. For most models, the number of cases classified as No Word is higher than those classified as Invalid Insert, indicating that models more frequently fail to include non-symbol characters outside the curly braces. In contrast, Gemini-2.0-Flash exhibits a larger number of failures categorized as Invalid Insert.

How does industry affect the diversity evaluations? Since model performance is influenced by industry differences (Mita et al., 2024), we conduct an industry-wise analysis of the diversity metrics. Figure 5 shows the results of the diversity metrics for each industry. The diversity metric values vary across industries. Furthermore, for pBLEURT, industries that exhibit relatively high values in the source data tend to also show relatively high values across the models. Appendix B provides detailed industry-wise results for each evaluation metric.

Impact of ad delivery constraints on diversity. We investigated diversity among generated templates satisfying the character-length and keyword-insertion format constraints evaluated by REG and KREG, respectively. We recomputed pBLEU and pBLEURT on this subset for comparison with the scores before this filtering step. Table 7 presents the results, with the Original row providing the existing-template baseline filtered under the same constraints. Across all models, the constraint-compliant subset has higher pBLEU and pBLEURT scores than the corresponding scores in Table 4, indicating lower surface diversity among the retained templates and greater semantic similarity among headlines instantiated from them. These differences reflect the properties of the outputs retained after filtering, rather than changes in model behavior caused by enforcing constraints during generation. Finally, Table 8 shows the original templates used in actual ad delivery and examples of generated templates.

	pBLEU (↓)	pBLEURT (↑)
Original	17.63± 8.43	**16.96**± 10.11
GPT-4o-mini	12.59± 4.45	12.48± 6.16
Gemini-2.0-Flash	**11.90**± 3.94	13.26± 6.15
CALM2-7B	14.20± 5.59	<u>13.96</u>± 7.94
CALM3-22B	<u>12.31</u>± 4.99	10.14± 5.29
Llama-3-ELYZA-JP-8B	13.59± 5.57	10.41± 5.80
Llama-2-ELYZA-JP-7B	13.78± 6.25	9.53± 6.57
Llama-2-ELYZA-JP-13B	16.01± 6.67	12.31± 7.34

Table 7: Surface diversity and semantic similarity for the subset satisfying the evaluated character-length and keyword-insertion format constraints. **Bold** is the highest value, <u>underline</u> is the second one. ± indicates the standard deviation.

Template source	Japanese	Translated English (gloss)
Original	{Keyword:看護師求人}をお探しなら	*If you are looking for {Keyword:nursing jobs}*
GPT-4o-mini	{keyword:おすすめの着圧下着}はこちら	*Find {keyword:recommended compression underwear} here*
Gemini-2.0-Flash	{Keyword:おすすめ}の{Keyword:北海道}ツアー	*{Keyword:Recommended} {Keyword:Hokkaido} tours*
CALM2-7B	{keyword:人気のアイテム}をチェック	*Check out {keyword:popular items}*
CALM3-22B	{Keyword:男性歓迎} 即入寮OK／求人	*{Keyword:Men welcome}; immediate move-in available / job openings*
Llama-3-ELYZA-JP-8B	取り付け簡単の{keyword:商品名}	*{keyword:Product name} that is easy to install*
Llama-2-ELYZA-JP-7B	{keyword:ぬいぐるみ}が豊富	*A wide selection of {keyword:stuffed toys}*
Llama-2-ELYZA-JP-13B	私が{keyword:ダイエット}に成功した	*I succeeded at {keyword:dieting}*

Table 8: Examples of original and model-generated templates with English glosses.

8 Conclusion

We proposed AdTEx, a benchmark framework for systematically evaluating LLM-driven advertisement template expansion. Automatic and manual evaluations show that current LLMs can increase surface diversity while generally satisfying character-length constraints for ad delivery. However, preserving semantic consistency, complying with formatting requirements, and maintaining content coherence remain challenging. Correlation and error analyses further reveal that generation quality depends on both the similarity of input headlines and the underlying LLM. We hope that AdTEx will facilitate future research on expanding practical advertising templates and contribute to more efficient real-world advertising operational workflows.

Limitations

Language. We conduct our experiments and analysis using Japanese advertisement text, and the findings are not intended to be directly generalized to all languages. However, we believe that the proposed template expansion framework can be readily adapted to other languages. Furthermore, as discussed in Section 2, Japanese has become a common testbed for advertisement generation research due to the maturity of its advertising ecosystem and the active research community. We hope that this study will provide a foundation for extending advertisement template expansion to other languages.

LLMs. We evaluated AdTEx using seven diverse LLMs, including both open- and closed-weight models, with varying parameter scales. Although evaluating additional LLMs would further broaden the benchmark, the primary contribution is the formulation of the advertisement template expansion task and the introduction of AdTEx as a benchmark framework. The selected models already exhibit diverse behaviors, enabling a comprehensive analysis of the task's characteristics and challenges. Therefore, we leave broader model coverage and future performance improvements to subsequent work.

Prompt. Prompt design is known to affect LLM outputs (Sakai et al., 2024b). Since this study focuses on generation through the instruction-following ability and inductive reasoning of LLMs, we explored only diverse template expansion from a single prompt. Although evaluation across diverse prompt templates would be desirable, we did not explore such variations because our primary focus is on proposing the task and its evaluation methodology.

Ethical Considerations

We carefully verified that the dataset contains no personally identifiable information, harmful content, or inappropriate material that would conflict with the data provider's guidelines. Therefore, we believe that this study fully complies with the ethical standards and considerations required for responsible research conduct. The data were provided by an advertising company under a formal data-sharing agreement. The agreement explicitly permits their use for this research, and all experiments were conducted in accordance with its terms. We used AI-assisted writing tools, such as DeepL and ChatGPT, for translation and grammatical refinement. However, all original scientific content, including the research ideas, methodology, experiments, analyses, and conclusions, was developed and written by the authors. The authors take full responsibility for the content of this paper. Based on the above, we believe that this study complies with the applicable ethical standards and that all

relevant ethical, licensing, and AI-use disclosures have been appropriately provided.

References

Zoë Abrams and Erik Vee. 2007. Personalized ad delivery when ads fatigue: an approximation algorithm. In *Proceedings of the 3rd International Conference on Internet and Network Economics*, WINE'07, page 535–540, Berlin, Heidelberg. Springer-Verlag.

Khalid Alnajjar and Hannu Toivonen. 2021. Computational generation of slogans. *Natural Language Engineering*, 27(5):575–607.

Yoichi Aoki, Soichiro Murakami, Ukyo Honda, and Akihiko Kato. 2025. Exploring the relationship between diversity and quality in ad text generation. *Preprint*, arXiv:2505.16418.

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, New York, NY, USA. Association for Computing Machinery.

Yanda Chen, Zihui Ren, Qixiang Gao, Jiale Chen, Si Chen, Xubin Li, Tiezheng Ge, and Bo Zheng. 2025. Ctr-driven ad text generation via online feedback preference optimization. *Preprint*, arXiv:2507.20227.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mbrs: A library for minimum Bayes risk decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–362, Miami, Florida, USA. Association for Computational Linguistics.

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, New York, NY, USA. Association for Computing Machinery.

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 & Data Mining*, KDD '19, page 2269–2277, New York, NY, USA. Association for Computing Machinery.

Go Inoue, Akihiko Kato, Masato Mita, Ukyo Honda, and Peinan Zhang. 2024. CAMERA^3: An evaluation dataset for controllable ad text generation in Japanese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2702–2707, Torino, Italia. ELRA and ICCL.

Daniel R. Jiang, Alex Nikulkov, Yu-Chia Chen, Yang Bai, and Zheqing Zhu. 2025. Improving generative ad text on facebook using reinforcement learning. *Preprint*, arXiv:2507.21983.

Yiping Jin, Akshay Bhatia, Dittaya Wanvarie, and Phu T. V. Le. 2023. Towards improving coherence and diversity of slogan generation. *Natural Language Engineering*, 29(2):254–286.

Hidetaka Kamigaito, Soichiro Murakami, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. 2024. Generating attractive ad text by facilitating the reuse of landing page expressions. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 597–608, Tokyo, Japan. Association for Computational Linguistics.

Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. 2021. An empirical study of generating texts for search engine advertising. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 255–262, Online. Association for Computational Linguistics.

Xiaoqiang Kang, Zimu Wang, Xiaobo Jin, Wei Wang, Kaizhu Huang, and Qiufeng Wang. 2025. Template-driven llm-paraphrased framework for tabular math word problem generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24303–24311.

Akihiko Kato, Masato Mita, Soichiro Murakami, Ukyo Honda, Sho Hoshino, and Peinan Zhang. 2024. Faithcamera: Construction of a faithful dataset for ad text generation. *Preprint*, arXiv:2410.03839.

Kentaro Kurihara, Masato Mita, Peinan Zhang, Shota Sasaki, Ryosuke Ishigami, and Naoaki Okazaki. 2025. Lctg bench: Llm controlled text generation benchmark. *Preprint*, arXiv:2501.15875.

Yukyung Lee, Soonwon Ka, Bokyung Son, Pilsung Kang, and Jaewook Kang. 2025. Navigating the path of writing: Outline-guided text generation with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 233–250, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexander Mafael, Sascha Raithel, Charles R. Taylor, and David W. Stewart. 2021. Measuring the role of uniqueness and consistency to develop effective advertising. *Journal of Advertising*, 50(4):494–504.

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 & Knowledge Management*, CIKM '20, page 2653–2660, New York, NY, USA. Association for Computing Machinery.

Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. 2024. Striking gold in advertising: Standardization and exploration of ad text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 955–972, Bangkok, Thailand. Association for Computational Linguistics.

Soichiro Murakami, Sho Hoshino, and Peinan Zhang. 2023. Natural language generation for advertising: A survey. *Preprint*, arXiv:2306.12719.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Cornelia Pechmann and David W. Stewart. 1988. Advertising repetition: A critical review of wearin and wearout. *Current Issues and Research in Advertising*, 11(1-2):285–329.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024a. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214, Bangkok, Thailand. Association for Computational Linguistics.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. Revisiting compositional generalization capability of large language models considering instruction following ability. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31219–31238, Vienna, Austria. Association for Computational Linguistics.

Yusuke Sakai, Adam Nohejl, Jiangnan Hang, Hidetaka Kamigaito, and Taro Watanabe. 2024b. Toward the evaluation of large language models considering score variance across instruction templates. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 499–529, Miami, Florida, US. Association for Computational Linguistics.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5719–5728. PMLR.

Stamatina Thomaidou, Konstantinos Leymonis, and Michalis Vazirgiannis. 2013a. Grammads: Keyword and ad creative generator for online advertising campaigns. In *Digital Enterprise Design and Management 2013*, pages 33–44, Berlin, Heidelberg. Springer Berlin Heidelberg.

Stamatina Thomaidou, Ismini Lourentzou, Panagiotis Katsivelis-Perakis, and Michalis Vazirgiannis. 2013b. Automated snippet generation for online advertising. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 1841–1844, New York, NY, USA. Association for Computing Machinery.

Chang Wang, Siyu Yan, Depeng Yuan, Yuqi Chen, Yanhua Huang, Yuanhang Zheng, Shuhao Li, Yinqi Zhang, Kedi Chen, Mingrui Zhu, and Ruiwen Xu. 2025. Beyond quality: Unlocking diversity in ad headline generation with large language models. *Preprint*, arXiv:2508.18739.

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 Conference on Knowledge Discovery & Data Mining*, KDD '21, page 3697–3707, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Peinan Zhang, Yusuke Sakai, Masato Mita, Hiroki Ouchi, and Taro Watanabe. 2025. AdTEC: A unified benchmark for evaluating text quality in search engine advertising. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7672–7691, Albuquerque, New Mexico. Association for Computational Linguistics.

Yan Zhuang and Yuan Sun. 2025. Pg-ce: A progressive generation dataset with constraint enhancement for controllable text generation. *Preprint*, arXiv:2509.17669.

A Examples of Prompt in Japanese

Figure 6 shows the original Japanese prompt for advertisement template expansion corresponding to Figure 3. Accordingly, Figure 7 shows the original Japanese prompt for template filtering corresponding to Figure 4.

Prompt for template expansion in Japanese

複数の具体的な広告文をもとに、共通構造を抽出した広告文テンプレートをいくつか作成してください。

テンプレートは以下の条件を厳守してください。

- テンプレートは広告文の構造のみを表し、実際の社名・サービス名・地名など具体的情報は含めないこと。
- 置換できない固定部分は全角文字で合計10文字以内とすること。
- 置換可能な部分には必ず{}を使用し、{}内に例として入る単語を1つ記載してください。
- 説明文や補足は一切記載せず、テンプレート本文のみ返すこと。
- 広告文が複数例与えられる場合、最も汎用性の高い共通の構造を1つにまとめてテンプレート化すること。

【出力形式】
期待する出力形式を参考にして、テンプレート文のみ返してください。

【広告文の実例】

- 子育て中の方へ【正社員の求人】
- 【コープ共済】妊娠中の方へ
- 【転職活動中の方へ】
- 床のワックスがけを検討中の方へ
- 子育て中の方でも働きやすい求人

【期待する出力形式】
{keyword:技術職}をお探しの方へ

Figure 6: Prompt for template expansion in Japanese.

Ad Template Extraction in Japanese

あなたは広告文の評価を行うアシスタントです。ユーザーから「templateから作成した広告文」が与えられます。

あなたが行うタスクは以下のとおりです：

- 広告文が「具体性があり、広告文として意味が成立するか」を判断する。
- 意味が成立しない・不自然・抽象的すぎる場合はFalseを返す。
- 広告文として不要な用語が含まれていたらFalseを返す。
- 意味が自然で広告文として利用できる場合のみTrueを返す

出力形式は以下に従うこと：

- True/False で判断する
- 必ず、True もしくは False のみを返すこと。

【**template**から作成した広告文】
{headline}

Figure 7: Prompt for filtering template in Japanese.

B Detailed Automatic Evaluation Results by Industry

We evaluated the impact of industry categories on automatic evaluation metrics. Detailed results are provided in Tables 9, 10, 11, and 12.

	GPT-4o	Gemini	CALM2-7B	CALM3-22B	ELYZA-JP-8B	ELYZA-JP-7B	ELYZA-JP-13B	Original
Academic interests	9.6 ± 3.52	11.39 ± 4.54	9.66 ± 4.02	7.32 ± 3.32	12.77 ± 6.04	7.64 ± 3.83	10.92 ± 5.49	9.68 ± 3.33
Automotives	8.03 ± 3.09	8.07 ± 2.89	6.64 ± 2.76	6.64 ± 3.09	8.78 ± 4.77	7.38 ± 3.27	7.04 ± 3.93	8.36 ± 3.44
Books and literature	18.1 ± 6.96	16.24 ± 5.77	16.29 ± 6.59	13.87 ± 5.83	11.86 ± 6.22	30.44 ± 13.34	8.47 ± 4.61	9.98 ± 4.15
Business and finance	8.06 ± 3.2	7.1 ± 2.6	6.67 ± 2.7	5.21 ± 2.52	7.78 ± 3.75	5.0 ± 2.68	7.59 ± 4.49	7.17 ± 3.1
Careers	9.37 ± 3.76	9.54 ± 3.7	8.42 ± 3.64	6.69 ± 3.15	7.35 ± 3.77	6.18 ± 3.35	7.64 ± 4.19	10.59 ± 4.57
Family and relationships	7.46 ± 3.45	6.85 ± 3.05	5.83 ± 2.43	6.84 ± 3.85	8.91 ± 4.08	10.0 ± 5.25	9.08 ± 4.2	9.84 ± 3.49
Food and drinks	12.2 ± 4.63	9.49 ± 3.65	9.37 ± 3.72	10.55 ± 4.66	14.57 ± 7.03	13.25 ± 5.31	11.27 ± 4.95	11.72 ± 4.24
Health	10.12 ± 4.17	9.8 ± 4.02	9.56 ± 4.31	10.5 ± 3.85	14.31 ± 6.12	11.91 ± 5.85	9.98 ± 5.49	13.45 ± 7.66
Healthy living	8.7 ± 3.74	8.51 ± 3.01	7.81 ± 3.69	6.72 ± 3.41	10.35 ± 6.56	7.33 ± 3.36	7.37 ± 5.02	14.78 ± 10.67
Hobbies and interests	9.11 ± 3.53	8.18 ± 3.52	7.43 ± 2.69	7.53 ± 3.82	11.7 ± 5.69	29.48 ± 9.3	17.08 ± 8.67	5.85 ± 2.64
Home and garden	8.61 ± 3.16	8.06 ± 2.61	7.54 ± 2.74	7.83 ± 4.5	8.76 ± 3.15	33.25 ± 13.47	7.1 ± 3.2	12.16 ± 5.35
Movies	11.45 ± 4.03	12.55 ± 5.96	11.52 ± 4.65	9.71 ± 4.29	20.86 ± 10.62	18.91 ± 11.49	15.84 ± 7.35	22.08 ± 4.16
Music and audio	10.24 ± 3.64	7.86 ± 2.91	9.6 ± 3.62	7.52 ± 3.29	10.06 ± 4.93	10.66 ± 5.09	8.92 ± 3.99	14.32 ± 4.45
News and politics	9.96 ± 4.08	10.18 ± 4.9	7.89 ± 3.57	6.39 ± 3.15	17.68 ± 6.54	22.52 ± 10.0	12.68 ± 8.9	28.53 ± 19.01
Personal finance	8.48 ± 3.19	9.84 ± 3.83	7.08 ± 2.76	5.9 ± 2.82	7.1 ± 3.66	6.04 ± 3.27	7.11 ± 4.09	8.49 ± 3.92
Pets	10.81 ± 4.19	11.29 ± 4.11	10.64 ± 3.85	9.78 ± 4.11	12.56 ± 6.25	15.27 ± 6.46	7.9 ± 3.75	14.14 ± 6.27
Pharmaceuticals, ...	11.36 ± 4.81	8.32 ± 3.78	7.25 ± 3.87	12.57 ± 8.22	15.67 ± 7.53	0.48 ± 0.07	5.36 ± 1.96	9.05 ± 3.46
Real estate	14.98 ± 5.66	18.6 ± 5.84	11.64 ± 4.53	9.38 ± 4.38	15.16 ± 6.76	11.08 ± 6.52	9.93 ± 4.3	19.74 ± 5.89
Shopping	11.42 ± 4.46	11.67 ± 4.1	9.44 ± 3.9	8.03 ± 3.71	10.92 ± 5.29	11.2 ± 5.08	11.39 ± 5.93	11.6 ± 5.0
Sports	9.05 ± 3.74	8.86 ± 3.31	9.24 ± 3.1	9.35 ± 3.75	12.87 ± 6.4	13.87 ± 9.32	10.4 ± 5.81	14.19 ± 5.44
Style and fashion	10.63 ± 4.59	9.05 ± 3.34	7.92 ± 3.22	6.77 ± 3.07	10.57 ± 5.33	10.95 ± 3.82	7.91 ± 4.6	18.3 ± 9.61
Technology and computing	11.05 ± 3.62	10.24 ± 3.68	9.73 ± 3.72	8.31 ± 3.89	8.9 ± 2.94	9.09 ± 3.48	10.89 ± 4.27	22.22 ± 11.84
Travel	8.62 ± 3.05	10.44 ± 3.61	7.6 ± 3.06	6.93 ± 3.17	7.87 ± 4.19	7.85 ± 3.26	8.13 ± 3.9	10.99 ± 4.49
Video gaming	11.54 ± 4.65	12.85 ± 5.07	9.54 ± 3.7	7.32 ± 3.41	9.34 ± 4.14	8.04 ± 3.85	6.86 ± 3.9	8.19 ± 3.41

Table 9: pBLEU results each industry. As a baseline, we refer to the values obtained from original templates. ±Values indicate the standard deviation.

	GPT-4o	Gemini	CALM2-7B	CALM3-22B	ELYZA-JP-8B	ELYZA-JP-7B	ELYZA-JP-13B	Original
Academic interests	9.6 ± 3.52	11.39 ± 4.54	9.66 ± 4.02	7.32 ± 3.32	12.77 ± 6.04	7.64 ± 3.83	10.92 ± 5.49	9.68 ± 3.33
Automotives	8.03 ± 3.09	8.07 ± 2.89	6.64 ± 2.76	6.64 ± 3.09	8.78 ± 4.77	7.38 ± 3.27	7.04 ± 3.93	8.36 ± 3.44
Books and literature	18.1 ± 6.96	16.24 ± 5.77	16.29 ± 6.59	13.87 ± 5.83	11.86 ± 6.22	30.44 ± 13.34	8.47 ± 4.61	9.98 ± 4.15
Business and finance	8.06 ± 3.2	7.1 ± 2.6	6.67 ± 2.7	5.21 ± 2.52	7.78 ± 3.75	5.0 ± 2.68	7.59 ± 4.49	7.17 ± 3.1
Careers	9.37 ± 3.76	9.54 ± 3.7	8.42 ± 3.64	6.69 ± 3.15	7.35 ± 3.77	6.18 ± 3.35	7.64 ± 4.19	10.59 ± 4.57
Family and relationships	7.46 ± 3.45	6.85 ± 3.05	5.83 ± 2.43	6.84 ± 3.85	8.91 ± 4.08	10.0 ± 5.25	9.08 ± 4.2	9.84 ± 3.49
Food and drinks	12.2 ± 4.63	9.49 ± 3.65	9.37 ± 3.72	10.55 ± 4.66	14.57 ± 7.03	13.25 ± 5.31	11.27 ± 4.95	11.72 ± 4.24
Health	10.12 ± 4.17	9.8 ± 4.02	9.56 ± 4.31	10.5 ± 3.85	14.31 ± 6.12	11.91 ± 5.85	9.98 ± 5.49	13.45 ± 7.66
Healthy living	8.7 ± 3.74	8.51 ± 3.01	7.81 ± 3.69	6.72 ± 3.41	10.35 ± 6.56	7.33 ± 3.36	7.37 ± 5.02	14.78 ± 10.67
Hobbies and interests	9.11 ± 3.53	8.18 ± 3.52	7.43 ± 2.69	7.53 ± 3.82	11.7 ± 5.69	29.48 ± 9.3	17.08 ± 8.67	5.85 ± 2.64
Home and garden	8.61 ± 3.16	8.06 ± 2.61	7.54 ± 2.74	7.83 ± 4.5	8.76 ± 3.15	33.25 ± 13.47	7.1 ± 3.2	12.16 ± 5.35
Movies	11.45 ± 4.03	12.55 ± 5.96	11.52 ± 4.65	9.71 ± 4.29	20.86 ± 10.62	18.91 ± 11.49	15.84 ± 7.35	22.08 ± 4.16
Music and audio	10.24 ± 3.64	7.86 ± 2.91	9.6 ± 3.62	7.52 ± 3.29	10.06 ± 4.93	10.66 ± 5.09	8.92 ± 3.99	14.32 ± 4.45
News and politics	9.96 ± 4.08	10.18 ± 4.9	7.89 ± 3.57	6.39 ± 3.15	17.68 ± 6.54	22.52 ± 10.0	12.68 ± 8.9	28.53 ± 19.01
Personal finance	8.48 ± 3.19	9.84 ± 3.83	7.08 ± 2.76	5.9 ± 2.82	7.1 ± 3.66	6.04 ± 3.27	7.11 ± 4.09	8.49 ± 3.92
Pets	10.81 ± 4.19	11.29 ± 4.11	10.64 ± 3.85	9.78 ± 4.11	12.56 ± 6.25	15.27 ± 6.46	7.9 ± 3.75	14.14 ± 6.27
Pharmaceuticals, ...	11.36 ± 4.81	8.32 ± 3.78	7.25 ± 3.87	12.57 ± 8.22	15.67 ± 7.53	0.48 ± 0.07	5.36 ± 1.96	9.05 ± 3.46
Real estate	14.98 ± 5.66	18.6 ± 5.84	11.64 ± 4.53	9.38 ± 4.38	15.16 ± 6.76	11.08 ± 6.52	9.93 ± 4.3	19.74 ± 5.89
Shopping	11.42 ± 4.46	11.67 ± 4.1	9.44 ± 3.9	8.03 ± 3.71	10.92 ± 5.29	11.2 ± 5.08	11.39 ± 5.93	11.6 ± 5.0
Sports	9.05 ± 3.74	8.86 ± 3.31	9.24 ± 3.1	9.35 ± 3.75	12.87 ± 6.4	13.87 ± 9.32	10.4 ± 5.81	14.19 ± 5.44
Style and fashion	10.63 ± 4.59	9.05 ± 3.34	7.92 ± 3.22	6.77 ± 3.07	10.57 ± 5.33	10.95 ± 3.82	7.91 ± 4.6	18.3 ± 9.61
Technology and computing	11.05 ± 3.62	10.24 ± 3.68	9.73 ± 3.72	8.31 ± 3.89	8.9 ± 2.94	9.09 ± 3.48	10.89 ± 4.27	22.22 ± 11.84
Travel	8.62 ± 3.05	10.44 ± 3.61	7.6 ± 3.06	6.93 ± 3.17	7.87 ± 4.19	7.85 ± 3.26	8.13 ± 3.9	10.99 ± 4.49
Video gaming	11.54 ± 4.65	12.85 ± 5.07	9.54 ± 3.7	7.32 ± 3.41	9.34 ± 4.14	8.04 ± 3.85	6.86 ± 3.9	8.19 ± 3.41

Table 10: pBLEURT results each industry. As a baseline, we refer to the values obtained from original templates. ±Values indicate the standard deviation.

	GPT-4o	Gemini	CALM2-7B	CALM3-22B	ELYZA-JP-8B	ELYZA-JP-7B	ELYZA-JP-13B	Original
Academic interests	89.05	93.61	85.49	85.20	91.23	89.13	25.38	100.00
Arts and culture	82.35	84.68	78.95	87.37	90.00	-	33.33	100.00
Automotives	92.01	96.19	77.55	81.81	84.98	84.50	50.00	99.81
Books and literature	90.72	86.10	86.06	88.55	93.15	92.59	42.17	100.00
Business and finance	88.75	94.37	82.34	84.32	86.16	86.24	37.43	100.00
Careers	89.75	95.91	86.82	83.38	87.71	88.77	28.38	100.00
Family and relationships	79.26	97.04	73.63	73.64	87.80	70.77	48.42	100.00
Food and drinks	88.94	90.21	85.00	83.51	92.90	89.78	78.64	100.00
Health	89.73	88.40	85.42	85.19	89.29	83.58	23.53	100.00
Healthy living	87.52	95.02	78.24	83.11	85.88	84.49	27.63	100.00
Hobbies and interests	82.11	82.86	77.27	78.22	94.44	100.00	66.67	100.00
Home and garden	84.53	85.81	77.32	79.51	83.39	85.71	71.25	100.00
Movies	85.33	87.82	58.78	78.18	83.92	76.32	32.65	100.00
Music and audio	88.64	89.59	83.95	86.60	92.12	100.00	65.00	100.00
News and politics	91.91	85.68	84.29	92.49	95.21	75.53	18.18	100.00
Personal finance	87.43	89.96	80.38	85.49	86.62	76.11	44.34	100.00
Pets	80.52	87.94	73.82	76.52	87.31	94.52	41.59	100.00
Pharmaceuticals, ...	88.16	80.86	61.54	90.79	85.11	100.00	0.00	100.00
Real estate	91.80	90.75	76.96	85.98	85.48	82.86	53.00	100.00
Shopping	87.18	93.51	81.32	82.80	87.45	86.60	30.00	100.00
Sports	89.73	86.42	70.00	87.65	83.72	100.00	33.33	100.00
Style and fashion	88.11	89.48	83.14	85.79	86.67	80.49	44.31	100.00
Technology and computing	83.70	86.35	81.52	82.75	75.43	90.35	33.33	100.00
Television	82.23	85.47	76.60	81.02	91.25	-	15.00	100.00
Travel	88.67	95.09	81.58	83.58	80.41	84.90	42.24	100.00
Video gaming	89.03	93.64	71.20	86.13	83.74	87.21	60.96	100.00

Table 11: REG results each industries. As a baseline, we refer to the values obtained from original templates.

	GPT-4o	Gemini	CALM2-7B	CALM3-22B	ELYZA-JP-8B	ELYZA-JP-7B	ELYZA-JP-13B	Original
Academic interests	71.37	24.50	50.89	72.36	57.59	85.18	56.06	83.17
Arts and culture	98.04	15.74	52.63	80.53	76.19	-	100.00	83.33
Automotives	80.16	18.74	68.63	82.10	68.55	88.12	64.73	92.22
Books and literature	80.50	6.70	39.86	76.25	63.67	92.59	51.74	73.94
Business and finance	74.96	19.40	60.52	79.61	65.74	91.30	64.71	81.34
Careers	72.89	16.71	54.33	73.97	62.01	84.99	56.49	85.03
Family and relationships	95.16	10.05	81.32	87.84	64.51	87.88	62.11	100.00
Food and drinks	91.70	14.64	66.43	86.77	78.42	92.81	62.14	95.92
Health	76.37	22.10	79.17	80.56	63.54	86.76	64.71	64.38
Healthy living	46.60	18.97	40.13	52.64	56.27	88.83	35.62	66.27
Hobbies and interests	77.89	6.67	59.09	69.31	78.95	80.00	0.00	39.66
Home and garden	94.57	23.30	41.04	86.14	71.53	100.00	72.50	97.20
Movies	74.48	16.71	70.23	89.61	71.15	73.68	95.92	73.33
Music and audio	95.84	16.27	28.40	90.08	73.81	80.00	51.67	90.91
News and politics	85.96	22.47	57.14	68.04	59.01	72.63	59.09	86.67
Personal finance	67.93	19.94	59.40	71.18	55.28	89.15	50.82	84.81
Pets	93.49	8.12	63.23	92.05	79.83	64.63	76.99	96.59
Pharmaceuticals, ...	75.00	4.32	65.38	76.32	66.00	100.00	50.00	58.82
Real estate	87.19	41.60	64.05	82.84	83.38	91.98	75.00	92.93
Shopping	75.73	16.01	62.16	78.08	68.65	91.29	66.56	91.70
Sports	67.81	8.09	52.50	66.26	51.16	25.00	66.67	81.48
Style and fashion	85.40	13.52	59.77	86.05	73.05	97.56	65.55	84.85
Technology and computing	72.56	15.65	61.96	75.60	63.59	90.43	41.94	93.84
Television	95.82	15.02	40.43	88.08	61.88	-	100.00	88.89
Travel	69.72	16.59	57.19	73.64	64.39	86.18	51.98	92.97
Video gaming	75.29	16.90	55.69	82.84	73.94	87.94	67.12	70.19

Table 12: KREG results for each industry. As a baseline, we refer to the values obtained from original templates.