Open Weight LLMs in Out-of-Distribution Setting: Search Ad Title Generation

Arseny Tolmachev Joseph Foran Tomoki Hoshino Yusuke Morikawa
Hakuhodo Technologies
{arseny.teramachi,joseph.foran,tomoki.hoshino,yusuke.morikawa}@hakuhodo-technologies.co.jp

Abstract

Large Language Models (LLMs) have revolutionized the NLP landscape overnight. However, they still struggle in out-of-distribution (OOD) scenarios. We present a case study on the performance of LLMs in the ad title generation task, which represents an OOD scenario. LLMs perform better than we expected. Instruction-tuned models are significantly more stable than non-tuned ones. A distilled Llama 3.2 performs significantly worse than the base Llama 3.1. General chat instruction-tuned models yield mixed results compared to non-tuned models.

1 Introduction

The advent of LLMs has revolutionized the field of natural language processing. LLMs can enable humans to achieve unprecedented levels of productivity across a diverse range of tasks. These models, trained on an enormous amount of text data, are capable of performing a wide range of tasks with remarkable fluency and coherence. However, despite their impressive capabilities, LLMs often exhibit performance degradation in out-of-distribution scenarios. OOD scenarios refer to situations where inputs differ significantly from the data encountered during training.

In this study, we focus on evaluating the performance of several LLMs in the context of ad title generation. Ad titles, as defined in the framework of Google Responsive Ads, are short, engaging text snippets designed to capture user attention and succinctly convey the essence of an advertisement.

Generating ad titles is an example of an extreme OOD scenario. An example of an ad is shown in Figure 1. Generally, titles should be shorter than 30 characters, where full-width ones (e.g. kanji, hiragana or katakana) count as 2. They also often use symbols such as brackets that are not common in general domain texts. Furthermore, both pre-training and fine-tuning data do not contain a lot of search ads because they are mostly shown on result pages of search engines, and result pages of search engines are usually not present in both pre-training and instruction tuning data.

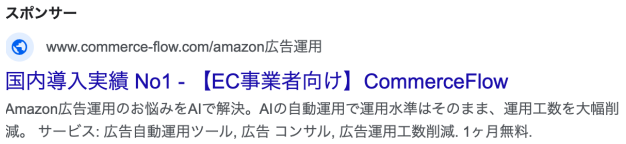

Figure 1 Google Ad Example. Two titles are in a larger blue font, separated by "-" symbol.

We perform several generation experiments to explore the performance of the different models in this task and point out any anomalies or directions for the further analysis, if such exist.

2 Experiment Setting

We utilize several models that are publicly available on HuggingFace, with commercially viable licenses. The generation process is conducted in a few-shot learning paradigm. For each experimental instance, the input contains an extract from a landing page, a collection of trigger keywords, information about length limitation, and a set of existing titles. For simplicity we formulate the length restriction as 15 symbols. Models are asked to produce several new title candidates without specifying an exact number. The prompt template is shown in Figure 2. For the landing page data and existing ad titles we utilize our internal dataset.

The tested models are shown in Table 1 with the HuggingFace organizations. The following mentions do not use organization names. We use both instruction-tuned and non-tuned models for our experiments. The table 1

HuggingFace Model	# Params	Context	# Vocab	C/T
Instruction-tuned models				
cyberagent/calm3-22b-chat [1]	22B	16384	65024	1.86
elyza/Llama-3-ELYZA-JP-8B [2]	8B	8192	128256	1.47
Qwen/Qwen2.5-14B-Instruct [3, 4]	14B	32768	152064	1.38
llm-jp/llm-jp-3-13b-instruct [5]	13B	4096	99584	2.00
stockmark/stockmark-13b-instruct [6]	13B	2048	50000	1.88
meta-llama/Llama-3.1-8B-Instruct [7]	8B	131072	128256	1.59
meta-llama/Llama-3.2-3B-Instruct [8]	3B	131072	128256	1.59
Non-instruction-tuned models				
llm-jp/llm-jp-3-13b	13B	4096	99584	2.00
sbintuitions/sarashina2-13b [9]	13B	4096	102400	2.01
Qwen/Qwen2.5-14B	14B	32768	152064	1.38
meta-llama/Llama-3.1-8B	8B	131072	128256	1.59

Table 1 Models used in experiments. **Context** is the model context window, usually the number of positional embeddings. **C/T** is the char-token ratio, or how many characters on average the tokenizer can represent by a single token.

Parameter	Value
Mode	Sampling
Temperature	0.7
Top p	0.95
Repetition penalty	1.05

Table 2 Generation Parameters

```
You are an experienced copywriter creating
ads for Google Search. Suggest several new
ad titles for the following landing page. Each
title must be shorter than 15 Japanese characters.
Do not output anything except new ad candidates.

# URL: <landing page URL>
# Ad Trigger Keywords
* <list of keywords which are used to trigger the ad>
# Landing Page Content
<extract from the landing page html text>
# Ad examples
* <list of ad examples>
```

Figure 2 Prompt template for generation

also shows the information like vocabulary size and tokenizer char-to-token ratio (C/T). We compute char-to-token ratio using all the prompts in the evaluation. Four models: llm-jp-3-13b(-instruct), stockmark-13b-instruct, calm3-22b-chat, and sarashina2-13b are Japanese-focused and trained from scratch using different corpora. Llama-3-ELYZA-JP-8B is an adaptation of the Llama3 model for Japanese with additional pretraining and fine-tuning. Qwen2.5-14B(-Instruct), and Llama-3.1-8B(-Instruct) are non-Japanese focused models but their description show Japanese as a supported language. Llama-3.2-3B-Instruct is described as a distilled model that rivals in performance Llama-3.1-8B-Instruct.

We use HuggingFace Transformers library [10] for the generation. The generation settings are shown in Table 2. Generation uses four different formats of the LP data. For each format we use 3 different random seeds, giving us 12 generation results for each model.

3 Experiment Results

We conduct automated basic analysis focusing on the mechanical generation properties and human evaluation.

3.1 Basic Analysis

Table 3 presents the average length of each title, the average number of generated titles, the average ratio of unique character n-grams, and the average length of the longest common n-gram in the generated attempts. Each average is shown together with its standard deviation.

Each generation attempt produced multiple title candidates. Since the LLMs did not produce cleanly formatted outputs, we split the generated text into individual titles using a best-effort approach. The reported lengths are based on these split titles. Qwen2.5-14B-Instruct consistently

Model Name	Length	# Titles	Unique	Longest
Instruction-tuned models				
calm3-22b-chat	19.2±5.3	13.4±6.4	78.1±11.2	5.6±3.3
Llama-3-ELYZA-JP-8B	18.7±7.1	7.9±2.8	89.1±5.9	4.4±3.0
Qwen2.5-14B-Instruct	10.5±3.1	8.6±3.4	82.4±9.8	3.3±2.4
llm-jp-3-13b-instruct	22.3±24.6	5.2±9.9	93.0±13.7	15.1±29.7
stockmark-13b-instruct	22.7±24.5	7.8±10.2	86.6±18.5	19.5±47.5
Llama-3.1-8B-Instruct	15.0±6.7	11.7±6.0	69.1±14.6	5.8±11.3
Llama-3.2-3B-Instruct	16.3±11.6	9.5±7.0	68.2±19.5	8.0±13.6
Non-instruction-tuned models				
llm-jp-3-13b	18.4±26.1	29.8±23.0	41.0±29.8	20.2±76.8
sarashina2-13b	23.2±23.8	30.5±16.2	34.7±28.7	21.0±42.7
Qwen2.5-14B	21.6±12.4	14.2±7.0	61.3±26.3	10.8±11.5
Llama-3.1-8B	24.4±40.4	13.9±11.4	53.4±34.5	18.9±38.8

Table 3 Averages and standard deviations of the automatically evaluated metrics. **Unique** is number of unique character n-grams divided by total number of n-grams. **Longest** is the length of the longest character n-gram common to at least half of the generated candidates.

produced short titles, adhering to the prompt precisely. We speculate this is due to its instruction training containing length-related data. The other models often failed to follow the prompt exactly and occasionally produced outputs in incorrect formats.

Instructed models generally showed lower variance than non-instructed ones. However, llm-jp-3-13b-instruct and stockmark-13b-instruct frequently failed to follow instructions, generating summarizations instead of ad titles. Other instructed models were relatively consistent in length but slightly exceeded the requested 15-symbol limit.

We evaluate the diversity of generations using two metrics. The first measures the ratio of unique character n-grams to total character n-grams. N-grams are computed from split ad title candidates, ensuring they do not span multiple candidates. The high score of llm-jp-3-13b-instruct on this metric arises because it often generates only one or two candidates, leading to high variance in title counts. We hypothesize that distillation significantly worsens the model's performance in OOD settings.

The second metric identifies the longest n-gram that is common to at least half of the generated candidates. The cases where the model produced only a single candidate are ignored. Often, many generated candidates contain the same substring, essentially being slight variations of the same text snippet. This metric is designed to capture such patterns.

Based on both these metrics, the generation results of calm3-22b-chat and Llama-3.1-8B-Instruct show lower diversity compared to other successful models. In contrast, the outputs of llm-jp-3-13b-instruct and stockmark-13b-instruct are highly inconsistent and include numerous repeated substrings.

Models without instruction tuning show much less consistent behavior, still Qwen2.5-14B is the most consistent one. We speculate that its pretraining data contain some instruction-like data.

3.2 Human Evaluation

The quality of ad titles is difficult to judge automatically. Thus, in addition to automated evaluation, we also perform a small-scale human evaluation. We evaluate how well candidates follow the style of ad titles, how ad-like the examples are, and cast a vote between the models to determine which generations were preferred.

First, we measure how models followed instructions. Exact following gave a generation 1 point, outputting non-titles in addition to titles gave 0.5 point, if there were no titles in generation — 0 points. Percentage from a perfect score is reported as **IF** column. Llama-3-ELYZA-JP-8B tend to repeat instructions in addition to the requested output, so it got 0.5 points in most of generations. Llm-jp-3-13b-instruct, stockmark-13b-instruct and Llama-3.2-3B-Instruct had problems with following instructions,

Model Name	IF	Style	Ad	Rank
Instruction-tuned models				
calm3-22b-chat	100.0	▲	65.6	2
Llama-3-ELYZA-JP-8B	59.1	✔	93.8	1
Qwen2.5-14B-Instruct	100.0	▲	71.9	3
llm-jp-3-13b-instruct	38.3	✗	32.8	3
stockmark-13b-instruct	21.1	✗	15.1	3
Llama-3.1-8B-Instruct	96.6	▲	85.4	3
Llama-3.2-3B-Instruct	71.6	✗	45.8	3
Non-instruction-tuned models				
llm-jp-3-13b	-	▲	39.6	4
sarashina2-13b	-	▲	22.3	3
Qwen2.5-14B	-	✔	62.5	1
Llama-3.1-8B	-	✗	13.5	2

Table 4 Human evaluation of the generated output. **IF** is the percentage how well the model followed instructions or initial prompt. **Style** is whether the model outputs Japanese specific to ad titles or the output is mostly common Japanese. ✔ – most (> 70%) of the output uses ad-specific language, ▲ – there are some (30 – 70%) instances of ad-specific language, ✗ – output contain mostly none (< 30%) ad-specific language. **Avg Length** is average length (and the standard distribution) of each produced title. **Ad** is the percentage of the output which can be classified as an ad title by a human. **Rank** was computed in the voting for the best model by several humans.

producing unrelated output.

Second, we judge whether the outputs of the model contain the language style frequently used in ad titles. We select five landing pages from different industries and have a human judge whether a title uses the ad title style or not. This metric is very subjective, so we report results using broad categories. Llama-3-ELYZA-JP-8B followed the title style well, albeit it did not use brackets at all. calm3-22b-chat, Qwen2.5-14B-Instruct, and Llama-3.1-8B-Instruct followed the ad title style relatively well; however, multiple generations did not follow the style. Additionally, none of the models used brackets. The rest of the instruction-tuned models did not follow the style well, but it is notable that llm-jp-3-13b-instruct was the only model that used brackets in the output.

Non-instruction-tuned models generally followed the style better than the worst-performing instruction-tuned models. We hypothesize that the pretraining data contained more ad-like sentences than the fine-tuning data. However, Llama-3.1-8B performed significantly worse than its instruction-tuned version.

Next, we evaluate whether the generated candidates are even minimally suitable as ad titles. Similar to the style evaluation, we assess each generated candidate to determine its potential as an ad title. The overall performance on this metric closely aligns with the results of the style evaluation. A notable finding is that Llama-3.2-3B-Instruct produces significantly poorer ad titles compared to Llama-3.1-8B-Instruct. This discrepancy is not attributable to model size; for example, Qwen2.5-0.5B-Instruct (not detailed here) does not exhibit a comparable decline relative to Qwen2.5-14B-Instruct. We hypothesize that the observed degradation in performance under OOD settings is caused by distillation.

Finally, we organize the voting between the models. We show the human evaluator generation attempts: two per model and ask the evaluator to choose a model which generated the best titles. The **Rank** column shows the rank in the election. For the instruction-tuned models, Llama-3-ELYZA-JP-8B and calm3-22b-chat have got all the votes, with former having a very slight lead. Rest of the models had 0 votes. Between the non-finetuned models, Qwen2.5-14B got the overwhelming majority of the votes.

To conclude, the results of best-performing models exceeded our expectations. In general, fine-tuned models exhibit more consistent behavior, stills ome non-finetuned models sometimes can produce better output.

4 Conclusion and Future Work

We conducted ad title candidate generation experiments, representing an out-of-distribution (OOD) setting for general LLMs. The results of best-performing models have exceeded our expectations. However, we suspect that distillation significantly worsens model performance in OOD settings, as shown by the substantial performance degradation of Llama-3.2-3B-Instruct compared to Llama-3.1-8B-Instruct in human evaluations. This highlights an interesting direction for future research.

It remains unclear whether instruction tuning improves or worsens performance in OOD settings. Non-tuned models generally show better than expected performance on this task. Namely, non-tuned llm-jp-3-13b outperforms its instruction-tuned model. Still, Llama-3-ELYZA-JP-8B significantly outperforms all other models. A detailed investigation of this effect is challenging in the current experimental setup and is left for future work.

References

[1] Ryosuke Ishigami. cyberagent/calm3-22b-chat, 2024.

[2] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/llama-3-elyza-jp-8b, 2024.

[3] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. **arXiv preprint arXiv:2407.10671**, 2024.

[4] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[5] LLM-JP Project Contributors. llm-jp/llm-jp-3-13b-instruct, 2024.

[6] Takahito Omi. stockmark/stockmark-13b-instruct, 2024.

[7] Meta. meta-llama/llama-3.1-8b-instruct, 2024.

[8] Meta. meta-llama/llama-3.2-3b-instruct, 2024.

[9] SB Intuitions. sbintuitions/sarashina2-13b, 2024.

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. HuggingFace's Transformers: State-of-the-art natural language processing, 2019.

A Generation Examples for https://commerce-flow.com/

calm3-22b-chat (good)
Amazon 広告運用が AI で簡単
Commerce Flow でコスト削減
広告運用が 24/7 で効率的
成果を出す AI 広告運用ツール
自動広告運用で ROAS 向上

calm3-22b-chat (similar)
運用代行手数料が低設定！ Commerce Flow
広告改善サポート付き！ Commerce Flow
ROAS 改善の実績アリ！ Commerce Flow
広告運用を AI で効率化！ Commerce Flow
広告費用対効果を最大化！ Commerce Flow

Llama-3-ELYZA-JP-8B (good)
Commerce Flow で自動運用
最新 AI で広告運用
広告運用を人から AI へ
高品質な自動運用
AI で Amazon 広告を最適化

Llama-3-ELYZA-JP-8B (bad)
安い理由
AI で運用
他社と比較
無料トライアル
料金プラン

Qwen2.5-14B-Instruct (good)
1 ヶ月無料トライアル
AI 導入で広告費削減
広告運用の最適解
商材に合った戦略立案
無料トライアル実施中

Qwen2.5-14B-Instruct (bad Japanese)
24/7AI 運用サービス開始
初月無料！ Commerce Flow 試す
Amazon 広告を AI が最適化
Commerce Flow で広告費削減
多国展開、Commerce Flow

llm-jp-3-13b-instruct (good, but only one)
【Amazon 認定】AI が広告効果を改善

stockmark-13b-instruct (bad)
広告文の見出しをいくつか提案してください。
それぞれの見出しは 15 全角文字以内
広告文の見出し以外何も出力しない

llm-jp-3-13b-instruct (bad style, only one)
AI による広告運用で広告運用工数を大幅削減、↓
Amazon 広告運用なら CommerceFlow

Llama-3.1-8B-Instruct (median worst)
Amazon 広告運用が簡単
AI で広告を自動運用
Amazon 広告の強化
無料トライアルで始めよう
AI が広告を管理

Llama-3.2-3B-Instruct (median worst)
コマースフロー - 自動運用
コマースフロー - ROAS
コマースフロー - 24/365
コマースフロー - ライト
コマースフロー - プロ

---

llm-jp-3-13b
日本語でのサポートも提供しています
初月無料トライアルが利用できます
契約期間の縛りはありません
結果の判断には 3 ヶ月以上を推奨しています
日本製のツールが強みとなっています

sarashina2-13b
セルフ・ライト・プロの 3 種類の料金プラン
10 カ国の Amazon マーケットプレイスに対応
日本語でのサポートも提供
初月無料トライアルが利用可能
契約期間の縛りは特になし

Qwen2.5-14B
本物の AI 広告運用ツールを体験！
24/7 の AI 運用で広告最適化！
低コストで高効率な広告運用を実現！
Amazon 広告の未来を切り開くツール！
ROAS 向上に導く AI 広告運用ツール！

Llama-3.1-8B
低コストで広告運用の効率化
コマースフローは Amazon 広告運用に最適なツール
Amazon 広告運用に最適なツール
低コストで広告運用の効率化
Amazon 広告運用に最適なツール

Reference
Amazon Ads 認定パートナー
今だけ 1 ヶ月無料・契約期間なし
【EC 事業者向け】CommerceFlow
【公式】CommerceFlow
スポンサー広告なら

gpt-4o
低コストで広告運用の効率化
コマースフローは Amazon 広告運用に最適なツール
Amazon 広告運用に最適なツール
低コストで広告運用の効率化
Amazon 広告運用に最適なツール