# Loss as a Data Introspection Method: Looking into Japanese Advertising Text Generation

Joseph Foran　Arseny Tolmachev
Hakuhodo Technologies
{joseph.foran,arseny.teramachi}@hakuhodo-technologies.co.jp

## Abstract

This paper presents a case study that examines the distribution of training loss values in a Japanese advertising text generation model. Using a LongT5 architecture, we analyze the characteristics of training examples that exhibit both high and low loss values. Our findings reveal several key patterns: low-loss examples often contain repetitive phrases and standardized advertising terminology, while high-loss examples tend to feature more complex grammatical structures and natural language patterns. We also identify potential issues in training data quality and discuss their implications for model performance. We find that measuring training loss per example in the training data is a useful diagnostic tool, for better understanding model characteristics.

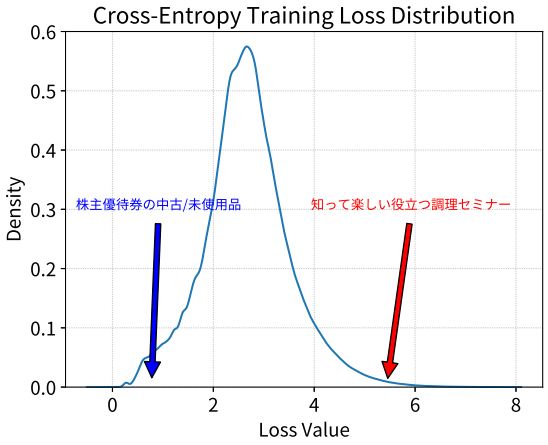


**Figure 1** A plot of a kernel density estimate of the loss values per training example for our model. We have annotated it with some examples from the training set and their position in the loss spectrum.

## 1 Introduction

Recent dramatic improvements in language models have resulted in significant improvements in many Natural Language Processing tasks. One such domain is the automatic generation of advertising text. As the share of online advertising, in particular search advertising, continues to grow within the overall advertising market, further automating this task is of increasing economic importance. In this work, we introduce a case study in which we examine the distribution of loss values across a dataset used to train a model that generates advertising text and present some findings that may be of interest to the community.

### 1.1 Advertising Text Generation

In this work, we focus on the generation of advertising text for search advertisements. These are advertisements that are displayed in connection to a user's search query, and the aim is to display advertisements that will be of relevance to the user based on keywords used in or related to the search query. Advertisers submit assets, consisting of headlines and description text, to an advertising platform. They associate keywords with such assets, and when such keywords are triggered an auction is conducted in order to decide whose advertisements will be displayed. The platform will display from the winning assets a combination of headlines and description texts, the composition of which are decided by a black-box algorithm. Because the combination that will be displayed is not known beforehand, the advertiser should ensure that any combination of headlines and descriptions is coherent.

The two categories of text that are to be generated by this model, headlines (sometimes referred to as titles), and descriptions, have differing characteristics. In particular, the length of the text and font size used when displaying on a search results page differ for these two types of text. Also, as the name suggests, headlines tend to be snappier, whereas the descriptions will be longer and have additional detail.

In general, Advertising Text Generation (ATG) models generate advertising text based on the contents of a Landing Page (LP) that contains details of the product / service being advertised. This task can be considered similar to summarization, though the strict length requirements and need to entice user's interest cause it to differ in significant ways from typical summarization tasks. Also, rather than a single summary, it is desirable to generate a large variety of texts based on the LP, each emphasizing different appeal points of the product or service being promoted.

To be effective, advertisements need to attract the user's interest, leading them to the LP with the aim of converting that interest into a sale. Effective advertisements convey information about the product/service that is of relevance to the user. Usually products and services have multiple features that appeal to different users. Generally the model should be able to generate many different texts that cover all these selling points. But while diversity in text generation is desirable, this is balanced by another important requirement in that they are truthful regardng what is on offer, as false claims can cause financial or reputational damage to the advertiser. For this reason, ATG models should also aim to be faithful to the contents of the LP which they use as input.

### 1.2 Data Rejuvenation

Our motivation for examining the training loss distribution was as a step in applying the Data Rejuvenation [1] method. This method identifies training examples that contribute less to the performance of the model, and replaces the target side of the training data with a synthetically generated example. Specifically, there are five general steps

1. Using all the original training data to train a model.
2. Identify each example's contribution to the model and divide the original training data into "effective" and "ineffective" sets.
3. Train a second model only on the "effective" set.
4. Use this model to generate pseudo-data using the source side of the "ineffective" dataset, thus "rejuvenating" these examples.
5. Use a combination of the "effective" dataset and the psuedo-data of the rejuvenated "ineffective" dataset to train a final model.

Jiao et. al [1] show that this methodology can improve the quality of Neural Machine Translation(NMT) models. Although this method was originally proposed for that particular task, in principle, it should be applicable in general to any sequence to sequence generation task. In this work, we use the idea of examining the training data loss spectrum to the aforementioned ATG task. That is we perform steps 1 and 2 above, and examine in more detail the distribution of training losses across all examples.

### 1.3 LongT5

The architecture of our generative model is the LongT5 model [2]. This is a Transformer [3] based model with an encoder-decoder structure. During training, the LP contents are fed into the encoder side, while the target advertising text is shown to the decoder, preceded by a prefix prompt token that indicates the target type i.e. headline or description. During inference the desired type of output can be controlled by setting this prefix prompt token.

## 2 Training Details

### 2.1 Training methodology

We pre-train a LongT5 model using Japanese portions of publicly available text corpora. We then fine-tune this model using an advertising text dataset. In this dataset, the source side of our examples contains text about the advertised good or service, while on the target side we use associated advertising texts.

## 3 Analysis of training loss distribution

After completion of training of our model, we measured the cross-entropy loss per example for all items in our training set, using the final model's parameters. In Figure 1 we show a density plot of this distribution, calculated using the Gaussian KDE algorithm [4] as implemented in SciPy [5].

It can be noticed that the distribution of cross-entropy loss is close to being unimodal. We examined examples at various points along the distribution of losses. In Table 1 (Headlines) and Table A2 (Descriptions) we provide examples from both the lowest, middle and highest deciles of cross-entropy loss.

When we examined examples from different parts of the spectrum, we found some characteristics that were of interest. Firstly, many low-loss targets were repeated frequently

| Low | Medium | High |
|---|---|---|
| ビルトイン食洗機が最大 70 %OFF | 初めてのお買い物で送料無料 | 脱毛が毎回違う店舗や人だけど… |
| 浴室乾燥機交換／最大 67 %OFF | 秋葉原で評判の歯医者 | あれあったっけが分かる冷蔵庫 |
| 株主優待券の中古/未使用品 | 折り畳み椅子 ロータイプ | 仮面様顔貌も脳の回復で改善 |
| 電気柵の中古/未使用品 | 電気殺虫器 | 室内でより確かな位置検出に |
| 格安新幹線＋宿泊パック比較 | 洗濯機で洗える羽毛布団 | 富栄養化を改善 |
| お役立ち資料無料配布中 | 破格の 100 枚 150 円から | 家の隙間から侵入する煙対策 |
| フィルター専門店だから種類豊富 | 発送ビニール袋 1 枚 14 円〜 | 出掛けすに家でスマホ副業 |
| 宅配買取で全国から高く買取ます | ライフコーチングを受ける | 台東嘉明湖 3 日團 免背公糧睡袋 |
| 無料サンプル最短当日で出荷 | 防犯カメラ 屋外 電池式 | 知って楽しい役立つ調理セミナー |
| 工事不要ですぐに使える | 車載ソーラーパネル | 海外で大人気のハンドメイド |
| 上海行きの格安フライト | 東京都の産業医をご紹介 | ブラウザ内で簡単に文字起こし |
| お手頃中古車や未使用車が多数 | 離れていても一体感を感じられる | 漢方茶とあなたの体質相性は？ |
| 和光市 ホテル | ペチコート ワンピース カップ付き | 頚椎を寝ながらストレッチする枕 |
| 花器 花瓶の人気アイテム | タフト シートカバー | 当日前日予約枠ご用意中 |
| 10 万円の少額から始められる | 団地間 4.5 畳 | 福祉ネイリストの需要が増えてる |
| 【法人向け】SMS 配信 1 通 5 円〜 | 長岡市で葬儀／小さなお葬式 | 食洗機取り付け業者大阪専門家 |
| バイク保険【安く抑えたい】方へ | スコップ収納ケースなら | 可愛い覆い袋にお納めしてご返骨 |
| 名古屋 激安 ホテル | 1 枚から格安で印刷、後払い OK | 高校で本格イラスト。体験授業も |
| 3L,4L,5L サイズの商品が豊富 | ホース径 変換 | 風呂の床だけリフォームが 1 日で |
| 【格安】月額￥2,900 から探せる | バイク売却 10 社一括査定 | 神奈川第 1 位の実績アリ |
| 年間販売台数 13 万台の販売実績 | 20 業種, 累計 2,500 万人以上が利用 | 土に落ちた成分は自然物に分解 |
| 新作が試せる！ 10/12 まで返品送料無料 | 屋外用ロールカーテン | 出しずらい高声も初回で上達 |
| 自転車 八王子の中古/未使用品 | 気になる使用感 | 事業承継で安定成長と雇用維持を |
| 銀行金利よりも高い利回り 7.0% | ネットワークの安定性向上 | 高精度工作機械メーカの歯研機 |

**Table 1** Headlines examples. These examples were sampled from the lower, middle and high deciles of the training loss spectrum.

in the data, with differing source examples. This indicates that there is possibly needless repetition of these examples which may harm the ability of the model to generate multiple diverse samples when given similar sources, as these examples become memorized. One characteristic of the ATG task that differs from many other NLP tasks is that novelty is especially required and so the introduction of too strong a training signal by repetitive targets, even if the sources for each target differs, should be avoided.

As was perhaps to be expected, among the high-level loss examples were found some examples that got through the training data cleaning / filtering process. In particular there were a few examples where the language was not Japanese and others where it was Japanese but there were misspellings or bad grammatical mistakes. These point out issues with the data collection process.

Another characteristic found was that many low loss examples followed well established patterns. For example, many low loss examples started with the phrase "【公式】" and were followed by a brand name [1)] which is commonly used in Japanese advertising text.

It was also noted that for both headlines and descriptions, the low loss examples tended to be more fragmented and to use phrase-like structures more. They tend to be declarative and snappy. On the other hand, high-loss examples tended to use more complete sentence structures and natural Japanese expressions. They also tend to use more varied grammatical structures, while there is a tendency for similar phrases to be repeated in the low-loss examples e.g 無料, 豊富なの品揃え。 Examples in the middle of the range were closer to everyday Japanese, having less commerical / retail vocabulary than the low loss examples, but at the same time tending towards having simpler vocabulary and kanji than the high-loss examples.

1） Examples containing trademarks or company names are excluded from this paper, so we do not provide specific examples.

The products referred to in these examples seem to be more everyday than the more specialised / niche products that the high-loss examples tend towards.

Focusing on Descriptions, low-loss examples tend to be list-like and categorical. This is likely also influenced by the nature of LPs which often contain such lists of features and products. The model in these cases has likely learned to assign high probability to phrases that it encounters in these contexts. High-loss examples tend to be more narrative and explanatory, while the mid-range examples fall somewhere in between, with more formal and direct language than the high-loss examples but with less incidents of repetitive, list-type structures.

## 4 Conclusions and Future work

Via this case study, we have shown that analysing the distribution of training data can help give good insights into the model, highlighting potential improvements to be made to the training data. In particular, examining the examples at the extremes of the training loss distribution provided insights, such as the presence of highly frequent target advertisements which may harm the model's ability to generate diverse outputs. We found that the linguistic characteristics of examples that the model found hard to learn differed from those it found relatively easy to learn. For future work, we plan to apply the rest of the Data Rejuvenation stages and investigate whether it leads to the improvement of the ability of our models to generate diverse and fluent advertising text.

## References


[1] Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. Data rejuvenation: Exploiting inactive training examples for neural machine translation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2255–2266. Association for Computational Linguistics, 2020.

[2] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. LongT5: Efficient text-to-text transformer for long sequences. In **Findings of the Association for Computational Linguistics: NAACL 2022**, pp. 724–736, Seattle, United States, 2022. Association for Computational Linguistics.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, NIPS'17, p. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[4] David W. Scott. **Multivariate Density Estimation: Theory, Practice, and Visualization**. John Wiley & Sons, New York, Chicester, 1992.

[5] SciPy Developers. **Gaussian KDE in SciPy**, 2023.

# A Description examples

| Loss level | Description |
|---|---|
| Low | 犬用品・猫用品・小動物用品・鳥用品・観賞魚用品など豊富な品揃え。注目アイテムもチェック。<br>接着・補修用品の豊富な品揃え：接着剤、粘着/養生テープ、潤滑油、補修材など。<br>エレクトロニクス部品の豊富な品揃え：制御、半導体部品から、センサ、コネクタまで。<br>スニーカー、サンダル、パンプスなど豊富な品揃え。靴のサイズ・幅・ヒールの高さから検索できる<br>電設用品の豊富な品揃え：スイッチ、コンセントプラグ、結束バンド、電設用資材など。<br>エレクトロニクス部品の豊富な品揃え：制御、半導体部品から、センサ、コネクタまで<br>安全・保護用品の豊富な品揃え：ヘルメット、防じんマスク、保護メガネ、軍手など。<br>メモリ 16GB、SSD 最大 2TB まで増加しても業界最安値の価格設定！<br>7 月 29 日金曜日 9 時から 7 月 31 日日曜日 23 時 59 分まで。大人気アイテム続々入荷中。<br>【低価格で高品質】デザイン、形、サイズ、素材、色、機能、価格帯など超充実！大型通販専門店<br>サプリから食品・コスメ・日用品まで。品質・原材料にこだわった商品が盛りだくさん！<br>毎月使ったデータ分だけ支払うお得なワンプラン。例えば 3GB までなら 980 円/月 (税込 1,078 円)。<br>お米・麺類・レトルト・調味料・スイーツなど豊富な品揃え。レビュー高評価商品をチェック。 |
| Medium | 150A 20k フランジの通販。配管部材やポンプ・ホースなど豊富に品揃え！ 入会金・年会費無料。<br>利用数 64 万人以上の経費精算システム。領収書のスキャンも簡単。電子帳簿保存法にも対応<br>一人ひとりのペースに合わせた指導だから早く成績が伸びる！ 小学生向け学習塾。<br>未公開の不動産情報も多数／無料会員登録で限定物件をご紹介！ お客様の声掲載中。<br>非公開講師も多数ご紹介。気になる講師の料金やスケジュール等ご相談ください！ 気になる講演料は<br>収入印紙や送付用封筒、謄本もまとめて注文できる。<br>福島のホテル&航空券。人気の宿泊施設や温泉宿を多数ご紹介。お得なパッケージツアー。<br>自社ローンだから審査が通り易い — 他店で審査に通らなかった方もご相談ください！<br>商品開発をはじめ、様々な取り組みも行っています。 豊富なラインナップ・ CM 情報・<br>天然 100 %のシルクで子供でも使えるほどの安心成分。 特別会員価格。野蚕のシルクパウダー。 |
| High | ラクラク自動集荷サービス。返送時は配送業者が返送用の宅配伝票を持参し引き取りにに伺います。<br>手の震え（振戦）やすくみ足は症状であり原因ではない。原因対策をお手伝いします。<br>IT ・ビジネス・産業の専門メディアに訪れる会員から、業種、役職クラス、企業規模などの属性指定<br>最大 32 名まで OK の掘りごたつ座敷。8 名× 4 卓あって、間仕切りで仕切ると個…<br>大自然のキャンパスで心身共に安定させ、勉強に部活に仲間と励み、最後は大学合格を目指します。<br>定額制で Web フォームが作り放題！いまこそあらゆる申し込み・問い合わせをクラウド化。<br>キレイめにまとまる Dress シリーズ。エレガントなヒールタイプの Womens シリーズなど種類豊富。<br>イラストアニメが分かりやすい！酵素サプリがきく理由徹底解説と人気商品との比較。<br>ページの挿入、削除、抽出、クロップ、順番変更、回転が簡単に実現できます。<br>水道管を使ったバー・ドアハンドルや、再利用素材を使った引出取手などのオリジナル金物が多数。<br>お土産は通販が一番！ WEB 注文でサッと手ぶら旅行へ。定番の和菓子はモチロン、梨ゼリーなど多数<br>シダー、フランキンセンス、ベチバーなどの特有のノートが楽しめるフレグランス。<br>ワタリガニ、カボス、豊後牛どれをとっても逸品揃いの大人気自治体。 |

**Table A2** Example Descriptions with indicated loss level. These examples include descriptions from both the lowest and highest decile of the loss spectrum.