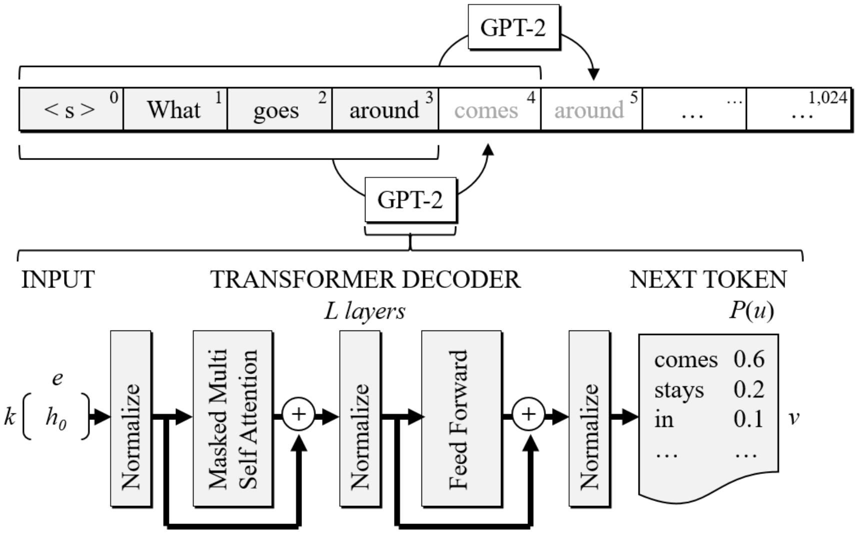
1 Technical Modeling and Validation Notes

1.1 GPT-2 Model Description

To provide the intuition behind transformer-based NLG models, we briefly illustrate the mechanics of the popular GPT-2 model. Given a sequence of tokens with context window size k, $U=(u_{-k},\dots,u_{-1})$, the objective of the autoregressive model GPT-2 is to accurately "predict" the next likely word[1] (Figure W1.1) by sampling from a probability distribution over its entire learned vocabulary (consisting of 50,257 tokens) conditional on the given word sequence and on a pre-trained neural network with parameters Θ. Model pre-training tries to maximize the likelihood in equation (W1) for an unsupervised corpus of words ($\mathcal{U}$) (Radford et al. 2018).

Figure W1.1: The GPT-2 Model[2]



[1] For ease of discussion, we describe the model in terms of "words." GPT-2 is derived using BPE (Byte Pair Encoding) and tokens (i.e., learned and encoded pieces of words).

[2] Visualization derived from Radford et al. (2018), and adapted to depict the updated GPT-2 architecture.

$$L_1(\mathcal{U}) = \sum_i logP(u_i|u_{i-k}, \dots, u_{i-1}; \Theta) \tag{W1}$$

$$h_0 = UW_e + W_p \tag{W2}$$

$$h_l = transformer\_block_{(h_{l-1})} \forall i \in [1, L] \tag{W3}$$

$$P(u) = softmax(h_L W_e^T) \tag{W4}$$

In essence, GPT-2 relies on word and given context meaning information to generate its output distribution over its vocabulary. More specifically, the data input consists of a matrix h_0 (W2), where the given word sequence U, word meaning information in terms of word embeddings W_e, and sequential word position information in terms of position embeddings W_p are combined. As illustrated in Figure W1.1, information from h_0 is extracted, transformed, added and normalized multiple times (to ease processing), and projected into the embedding space e by L layers of decoder transformer blocks (W3). This information includes the extent of putting attention on a given word sequence using multi-headed self attention (Vaswani et al. 2017), and high dimensional hidden language states to shift the focus in the embedding space e to recreate natural word sequences from position-wise feed forward neural networks. The output of the final block h_L projects all this information into the embedding space and is multiplied with GPT-2's original (unconditional) transposed word embeddings matrix W_e^T to assess which word from the GPT-2 vocabulary best matches the information contained in h_L (W4). The multiplication of h_L and W_e^T can be thought of as a similarity or matching between the embedding space distribution of the output of h_L (containing meaning, position, attention, and hidden language states information) and the unconditional embedding space distribution of each respective vocabulary word. More similarity of a vocabulary word in terms of its embedding to h_L will result in a higher probability in GPT-2's output distribution. GPT-2 then obtains a probability distribution over its vocabulary $P(u)$ (W4) and can sample the upcoming word in the sequence from the most likely words in $P(u)$.

Using the above procedure, GPT-2 learned and stored word probabilities for given word sequences represented in its 345 million parameters (including embeddings, attention weight matrices and Θ) using 8 million English text documents with a broad topical variety. Neural network parameters Θ were first initialized and then trained on batches of 512 sequences. The loss function refers to the language modeling cross entropy loss, where 1 is assigned to the word that appears next (u_i) in the training sequence (e.g., “comes” in Figure W1.1), and 0 to all other words in GPT-2’s vocabulary, and compare the log transformed GPT-2 softmaxed output probability value P_u for that respective word to appear next. A loss of 0 means the GPT-2 prediction was in perfect accordance with the actual next word (i.e., 1), the higher the deviation of the GPT-2 prediction (e.g., 1-0.6 = 0.4 for “comes” in Figure W1.1) to the actual word, the more the loss value increases. During training, GPT-2 performs this process on batches and mini-batches of several sequences before updating Θ.

1.2 Description of Essential Software Features

For our content generation method to work “on the fly,” a number of features have been incorporated at each stage of our algorithm for it to work automatically and reliably for any specified keyword. We summarize the most essential high-level features in Table W1.2.

Table W1.2: Developed Content Writing Machine Software Features

Step	Feature	Feature description
Ranking & links crawling	Crawler status updates	The crawler provides live-status updates
	Human behavior simulation	The crawler simulates human behavior to not get blocked by the search engine
	Organic links detection	Organic links are automatically detected and other links (e.g., paid ads) are discarded
	Duplicate entries detection	Duplicate ranking entries (e.g., featured content snippet) are detected and discarded
	Link correction (abbreviations, prefixes, etc.)	Abbreviated organic link prefixes are detected and corrected
	Data handling & saving	The crawler handles and saves the data in a structured way
Content scraping	Scraper status updates	The scraper provides live-status updates
	Local user client simulation	The scraper simulates an actual user client including a local OS, a local browser and a record of cookies
	Enhanced SSL protocols	The scraper uses enhanced SSL-protocols
	Website main content recognition	The scraper analyzes the HTML-files as well as common text patterns to find the main content on the webpage (i.e., discarding content of footers, main menus, etc.)
	Code cleanup	The scraper detects HTML, CSS & Java script code and cleans the main content from it
	Text cleanup	The scraper detects unwanted text snippets and patterns (i.e., big empty spaces, unintended line breaks, citing, remarks, etc.) and cleans the main content from it
	Server failure messages cleanup	The scraper universally detects server messages (e.g., “you are not allowed to crawl this website”, “502 server error”, etc.) and discards these
	Multi-redundancy	The scraper consists of multiple safety lines for full automation including error detection and handling, timeouts on jobs, etc.
	Auto-output	The scraper automatically outputs a txt-file that includes all information for retraining, including the main content, the targeted main keyword, and automatically generated special tags
Fine-tuning & content generation	Dynamic retraining & content generation	Our method performs a dynamic fine-tuning (i.e., several retraining, model checkpoint-saving and text generation steps where in each step, the retraining is continued to cover the full spectrum of model fine-tuning and fitting on the data)
	Fallback model	Our method uses the base model as fallback (i.e., if the retraining material is corrupt, it still generates and provides well written texts from an early retraining phase)
	H1, auto seed-word, and auto tagging	Our method takes the specified main keyword automatically as text headline and as seed-word for text generation for increased text consistency and topic focus
	Model checkpoints & file saving	Our method automatically performs several model-checkpoint savings and outputs the generated content in structured txt-files
Content selection & output	Auto data handling	The generated content is automatically cleaned and handled
	Auto quality score calculation	The quality score for text selection is automatically calculated
	Redundancy (error handling)	Errors are detected, outputted and appropriately handled
	Intuitive ordered list output for humans	An intuitive annotated output in the form of an ordered list of suggested generated and selected content is provided to a human reviewer

1.3 Applied Uniqueness, Naturality & Readability Measures

Without loss of generality, the quality score we present in the article could be adapted to incorporate other linguistic components. The software tool employed in our empirical application studies implements the components content uniqueness (s_d), naturality similarity (s_n) and readability similarity (s_r) as follows:

Uniqueness measurement (s_d). For our quality score (qs_g), we derive a uniqueness measure (s_d) to assess if the content is sufficiently unique for the search engine. In addition to the definitions around formula (3) in the main manuscript, we apply a critical value (s_{cv}) to ensure that the generated content is sufficiently unique based on the length of the keyword (kw) and parameter b.

$$s_{cv} = (100 - (100/(kw + 1)^b))/100 \tag{W5}$$

By implementing this non-compensatory filtering rule we ascertain that content that fails to achieve this minimum level of uniqueness is discarded from further content selection. The value b determines the factor of increasing conservativeness the larger the n-gram size (kw+1), as repeating small sized n-grams is less of a concern than repeating large sized n-grams (W5). In our setup, we set b to 1.1 after an evaluation phase in which we look at a) the machine output, b) acceptable duplicate rates in human content impressions, and c) content retaining rates for the whole range of common n-gram sizes. For example, that means that with an n-gram size of 3, s_{cv} ~.70 (i.e., 70% unique), an n-gram size of 5, s_{cv} ~.82 (i.e., 82% unique), and an n-gram-size of 7, s_{cv} ~.88 (i.e., 88% unique).

Naturality similarity measures (s_n). To quantify the naturality similarity between the generated content and the top ranked search results, we applied 12 linguistic measures which assess the lexical richness and composition of a text using the R package languageR.

Specifically, we use the following measures: tokens, types, hapax legomena, dis legomena, tris legomena, Yule’s K, Zipf’s R, Type-Token-Ratio, Herdan’s C, Guiraud’s R, Sichel’s S, Lognormal. More information on the precise meaning, practical examples and literature sources can be found in Baayen and Shafaei-Bajestan (2019).

Readability similarity measures (s_r). For the readability similarity measure, we applied 46 pre-existing measures of readability contained in the R package quanteda (see Benoit et al. 2018). We make use of the following measures: ARI, Bormuth.MC, Bormuth.GP, Coleman, Coleman.C2, Coleman.Liau.ECP, Dale.Chall, Dale.Chall.PSK, Danielson.Bryan, Dickes.Steiwer, DRP, ELF, Farr.Jenkins.Paterson, Flesch, Flesch.PSK, Flesch.Kincaid, FOG, FOG.PSK, FOG.NRI, FORCAST, FORCAST.RGL, Fucks, Linsear.Write, nWS, nWS.2, nWS.3, nWS.4, RIX, Scrabble, SMOG, SMOG.C, Spache, Spache.old, Strain, Traenkle.Bailer, W, St, C, Sy, W3Sy, W2Sy, W_1Sy, W6C, W7C, Wlt3Sy, W_wl.Dale.Chall. More information on the precise meaning, calculation and literature sources can be found in Benoit et al. (2018).

1.4 External Validation of Method Assumptions and Quality Score

Before using our method in a field application, we empirically test and confirm that the highest-ranking websites in the search engine indeed score highest in terms of our developed quality score components. For this task, we used around 8,500 relevant keywords and about 1.42 million ranked websites from 4 main industry sectors and 36 specific industries: details on the distribution of these keywords across industries and the number of scraped website content are reported in Table W1.4.1, columns (1) and (2). Using Wilcoxon rank sum group comparison tests, Table W1.4.2 illustrates that the poorer the search engine ranking, the lower the quality scores compared to the top 10 ranked content tends to be for all quality score components with the exception for content uniqueness (s_d).

Recall that the content uniqueness score s_d compares a given piece of content to the top 10 search results. The observed pattern in Table W1.4.2 suggests that a given result from outside of the top 10 set is more unique compared to the top 10 set than a result from the top 10 set is unique compared to other top 10 results. As noted in the manuscript, this may arise because the top 10 ranked websites consistently reflect similar topics. A contributor to the similarity among the top 10 ranked websites are the topics and keywords discussed in the content, as reflected by the higher scores for (s_a and s_k, respectively) compared to lower ranked websites. In assessing uniqueness relative to the content of the top 10 ranked websites, as these sites tend to cover similar topics and keywords compared to lower ranked sites, it is not surprising to find that the uniqueness score among the top 10 ranked websites is lower than the uniqueness score associated with lower ranked websites. Thus, we can ascertain that fine-tuning on the top 10 ranked websites' content will produce the most optimal content, and supporting the use of our quality score as a measure of content optimality. We aggregate the search engine results into groups (i.e., top 10, search engine ranks 11-20, search engine ranks 21-99, search engine ranks 100-200) to summarize the results.

Table W1.4.1: Empirical Setup for Validating Method and Quality Score Assumptions

Industry Sector	**Industry**	**(1) Number of Keywords**	**(2) Number of Scraped Rankings & Websites**	**(3) Number of Selected Keywords**	**(4) Number of Generated Texts**
I.	Coal Mining	100	14,678	5	5,000
	Forestry	501	87,537	9	9,000
	Grazing	100	18,021	10	10,000
	Hunting	100	17,621	7	7,000
	Fishing	500	77,210	10	10,000
	Quarrying	176	18,448	8	8,000
II.	Automobile production	270	42,303	10	10,000
	Textile production	150	26,960	9	9,000
	Chemical engineering	230	43,288	8	8,000
	Aerospace production	250	57,149	10	10,000
	Energy utilities	150	29,767	10	10,000
	Breweries & bottlers	150	30,691	9	9,000
	Construction	150	21,757	7	7,000
	Ship building	70	14,058	9	9,000
	Jewelries	245	45,097	9	9,000
III.	Retailing	150	27,717	9	9,000
	Transportation	450	60,222	9	9,000
	Restaurants	230	32,539	9	9,000
	Clerical service	300	49,188	9	9,000
	Mass media	300	39,784	9	9,000
	Tourism	300	41,174	10	10,000
	Insurance	150	27,581	10	10,000
	Banking	270	44,007	9	9,000
	Healthcare	150	30,478	10	10,000
	Law	230	43,717	9	9,000
	IT service	324	50,670	19	19,000
	Art & galleries	150	27,167	9	9,000
	Cafes	230	35,382	9	9,000
	Grocery stores	500	80,814	10	10,000
	Media agencies	150	29,180	10	10,000
IV.	Government	300	50,074	9	9,000
	University	349	54,775	11	11,000
	Culture	300	57,704	9	9,000
	Libraries	100	15,715	9	9,000
	Research	100	9,938	10	10,000
	Education	278	62,518	10	10,000

Table W1.4.2: External Validation of Method Assumptions Statistics

Industry Sector	Ranks of Content Compared to Top 10	Topic (s_a)[1]	Keywords (s_k)[1]	Uniqueness (s_d)[1]	Readability similarity (s_r)[1]	Naturality similarity (s_n)[1]
I.	Top 10	.27 (.16)	.23 (.23)	.93 (.22)	.74 (.57)	.75 (.50)
	11 - 20	.23 (.17)**	.18 (.23)**	.96 (.11)**	.70 (.62)**	.58 (.58)**
	21 - 99	.18 (.15)**	.13 (.20)**	.96 (.09)**	.65 (.55)**	.58 (.58)**
	100 - 200	.15 (.15)**	.09 (.20)**	.97 (.09)**	.70 (.62)**	.67 (.58)**
II.	Top 10	.31 (.17)	.26 (.22)	.95 (.15)	.70 (.57)	.67 (.50)
	11 – 20	.25 (.16)**	.20 (.22)**	.97 (.09)**	.62 (.62)**	.58 (.50)**
	21 - 99	.22 (.17)**	.16 (21)**	.97 (.08)**	.59 (.59)**	.58 (.50)**
	100 - 200	.17 (.15)**	.11 (.21)**	.97 (.07)**	.57 (.57)**	.50 (.50)**
III.	Top 10	.35 (.22)	.31 (.30)	.94 (.17)	.72 (.60)	.75 (.50)
	11 - 20	.29 (.21)**	.25 (.29)**	.96 (.10)**	.70 (.60)**	.67 (.58)**
	21 - 99	.23 (.20)**	.17 (.26)**	.97 (.08)**	.64 (.60)**	.58 (.58)**
	100 - 200	.18 (.17)**	.10 (.22)**	.98 (.06)**	.57 (.62)**	.50 (.58)**
IV.	Top 10	.31 (.20)	.26 (.27)	.95 (.10)	.72 (.60)	.62 (.58)
	11 - 20	.27 (.20)**	.21 (.25)**	.97 (.08)**	.68 (.57)**	.58 (.58)**
	21 - 99	.22 (.19)**	.14 (.21)**	.97 (.07)**	.62 (.60)**	.57 (.58)**
	100 - 200	.16 (.16)**	.07 (.18)**	.97 (.06)**	.62 (.59)**	.42 (.67)**

[1] Reported numbers are group medians and IQRs in parentheses. Statistical significance codes come from Wilcoxon rank sum 2-group comparison tests between top 10 ranked websites and the content with specific rankings as stated in column 2; statistical significance codes (one-tailed): *0.05 level, **0.01 level; assumptions (e.g., non-normality of data) for all Wilcoxon rank-sum 2-group comparison tests are confirmed.

The results of Table W1.4.2 are consistent for smaller sets of search engine rankings (which correspond to a given page of search engine results) for the single exemplary industry sector III (Table W1.4.3) and for specific industries (e.g., tourism) (Table W1.4.4). The observed patterns are consistent across industries and the tests for significance are insensitive to industry sector aggregations and/or data-groupings; more detailed industry-specific results are available from the authors upon request.

Table W1.4.3: External Validation of Method Assumptions Statistics for Industry Sector III

Industry Sector	**Ranks of Content Compared to Top 10**	**Topic (s_a)[1]**	**Keywords (s_k)[1]**	**Uniqueness (s_d)[1]**	**Readability similarity (s_r)[1]**	**Naturality similarity (s_n)[1]**
III.	Top 10	**.35 (.22)**	**.31 (.30)**	**.94 (.17)**	**.72 (.60)**	**.75 (.50)**
	11 - 20	.29 (.21)**	.25 (.29)**	.96 (.10)**	.70 (.60)**	.67 (.58)**
	21 - 30	.27 (.20)**	.22 (.27)**	.96 (.09)**	.66 (.60)**	.58 (.50)**
	31 - 40	.25 (.20)**	.20 (.27)**	.97 (.09)**	.66 (.62)**	.58 (.58)**
	41 - 50	.24 (.20)**	.19 (.26)**	.97 (.08)**	.63 (.59)**	.58 (.58)**
	51 - 60	.23 (.19)**	.17 (.25)**	.97 (.08)**	.64 (.57)**	.58 (.58)**
	61 - 70	.23 (.19)**	.16 (.25)**	.97 (.08)**	.63 (.62)**	.58 (.67)**
	71 - 80	.22 (.19)**	.15 (.25)**	.97 (.08)**	.66 (.62)**	.58 (.67)**
	81 - 90	.21 (.18)**	.15 (.24)**	.97 (.08)**	.64 (.64)**	.50 (.58)**
	91 - 100	.19 (.18)**	.12 (.23)**	.97 (.08)**	.62 (.62)**	.50 (.58)**
	101 - 110	.19 (.19)**	.12 (.24)**	.97 (.07)**	.61 (.62)**	.50 (.58)**
	111 - 120	.19 (.18)**	.11 (.23)**	.98 (.07)**	.62 (.59)**	.50 (.58)**
	121 - 130	.20 (.18)**	.12 (.24)**	.98 (.07)**	.57 (.55)**	.50 (.58)**
	131 - 140	.18 (.19)**	.11 (.25)**	.97 (.08)**	.60 (.55)**	.50 (.58)**
	141 - 150	.17 (.16)**	.10 (.20)**	.97 (.07)**	.62 (.59)**	.50 (.58)**
	151 - 160	.16 (.16)**	.09 (.20)**	.97 (.07)**	.53 (.57)**	.42 (.67)**
	161 - 170	.16 (.16)**	.09 (.21)**	.97 (.08)**	.57 (.62)**	.50 (.58)**
	171 - 180	.16 (.15)**	.08 (.19)**	.97 (.07)**	.55 (.56)**	.50 (.58)**
	181 - 190	.16 (.15)**	.08 (.19)**	.97 (.07)**	.61 (.62)**	.58 (.58)**
	191 - 200	.15 (.14)**	.07 (.17)**	.97 (.07)**	.49 (.55)**	.50 (.58)**

[1] Reported numbers are group medians and IQRs in parentheses. Statistical significance codes come from Wilcoxon rank sum 2-group comparison tests between top 10 ranked websites and the content with specific rankings as stated in column 2; statistical significance codes (one-tailed): *0.05 level, **0.01 level; assumptions (e.g., non-normality of data) for all Wilcoxon rank-sum 2-group comparison tests are confirmed.

Table W1.4.4: External Validation of Method Assumptions for the Tourism Sector

Industry	**Ranks of Content Compared to Top 10**	**Topic (s_a)[1]**	**Keywords (s_k)[1]**	**Uniqueness (s_d)[1]**	**Readability similarity (s_r)[1]**	**Naturality similarity (s_n)[1]**
Tourism	Top 10	**.37 (.32)**	**.36 (.40)**	**.88 (.27)**	**.63 (.60)**	**.58 (.58)**
	11 - 20	.32 (.21)**	.30 (.29)*	.95 (.18)**	.60 (.53)	.50 (.63)*
	21 - 30	.30 (.22)**	.26 (.28)**	.95 (.09)**	.48 (.54)**	.42 (.58)**
	31 - 40	.28 (.32)**	.26 (.26)**	.95 (.11)**	.42 (.48)**	.42 (.50)**
	41 - 50	.28 (.20)**	.26 (.27)**	.96 (.08)**	.49 (.53)**	.41 (.50)**
	51 - 60	.25 (.16)**	.21 (.24)**	.97 (.08)**	.53 (.51)**	.42 (.58)**
	61 - 70	.25 (.19)**	.23 (.27)**	.95 (.11)**	.43 (.52)**	.42 (.50)**
	71 - 80	.23 (.18)**	.18 (.27)**	.97 (.08)**	.43 (53)**	.33 (.50)**
	81 - 90	.20 (.21)**	.14 (.25)**	.97 (.07)**	.51 (.55)**	.33 (.50)**
	91 - 100	.23 (.21)**	.17 (.28)**	.96 (.09)**	.43 (.59)**	.33 (.48)**
	101 - 110	.19 (.19)**	.15 (.26)**	.97 (.10)**	.43 (.57)**	.33 (.58)**
	111 - 120	.16 (.17)**	.10 (.22)**	.97 (.08)**	.34 (.52)**	.25 (.35)**
	121 - 130	.16 (.18)**	.07 (.22)**	.96 (.11)**	.32 (.51)**	.33 (.46)**
	131 - 140	.14 (.15)**	.08 (.20)**	.97 (.07)**	.36 (.49)**	.42 (.46)**
	141 - 150	.17 (.19)**	.13 (.23)**	.97 (.10)**	.33 (.55)**	.29 (.56)**
	151 - 160	.12 (.10)**	.06 (.14)**	.97 (.07)**	.46 (.40)**	.25 (.25)**
	161 - 170	.10 (.11)**	.04 (.10)**	.96 (.09)**	.33 (.45)**	.38 (.50)**
	171 - 180	.16 (.12)**	.10 (.16)**	.97 (.11)**	.51 (.49)**	.33 (.48)**
	181 - 190	.12 (.19)**	.04 (.18)**	.97 (.03)**	.46 (.62)**	.25 (.50)**
	191 - 200	.13 (.11)**	.10 (.14)**	.98 (.07)**	.59 (.48)	.50 (.58)

[1] Reported numbers are group medians and IQRs in parentheses. Statistical significance codes come from Wilcoxon rank sum 2-group comparison tests between top 10 ranked websites and the content with specific rankings as stated in column 2; statistical significance codes (one-tailed): *0.05 level, **0.01 level; assumptions (e.g., non-normality of data) for all Wilcoxon rank-sum 2-group comparison tests are confirmed.

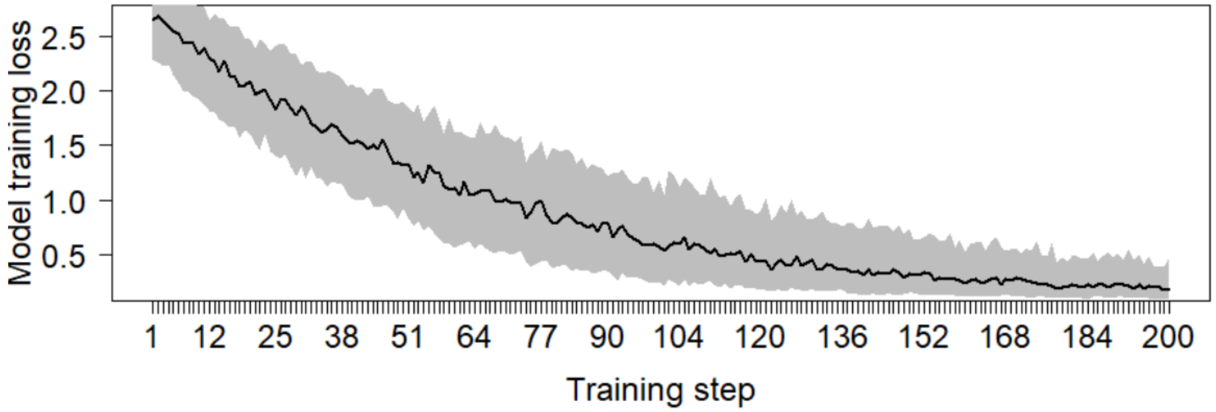
1.5 Validation of Method Fine-Tuning Process

For our experiments, we fine-tune our model for 200 training steps for each keyword, generating 100 pieces of content at each 20th training step which resulted in 1,000 generated texts per focal keyword, of which our method then selected the best scoring pieces of content using the proposed quality score metric. Similar to the approach taken by Liu and Toubia (2018), based on prior literature and on several test runs, we set the hyper-parameters top_*k* = 40, and temperature = 0.7 (which effectively regulates the randomness in GPT-2's sampling process and output content). Next, we show that fine-tuning for 200 training steps is sufficient and examine factors that determine at which training step our proposed method selects the most optimal content.

Figure W1.5.1 illustrates the increasing capability of the model to accurately predict words given prior word sequences over the 200 model training steps using the median (black line) and IQR (grey area) of the Loss measure (Radford et al. 2018) over all keyword trainings for the experiments presented in the manuscript. While model fit is consistently improving, Figure W1.5.2 shows that the most optimal content on the basis of the quality score commonly comes from mid training steps (between 60 and 160), while an extremely low and an extremely high amount of training steps entail a lower probability to produce the optimal content. Thus, using 200 training steps for fine-tuning is sufficient.

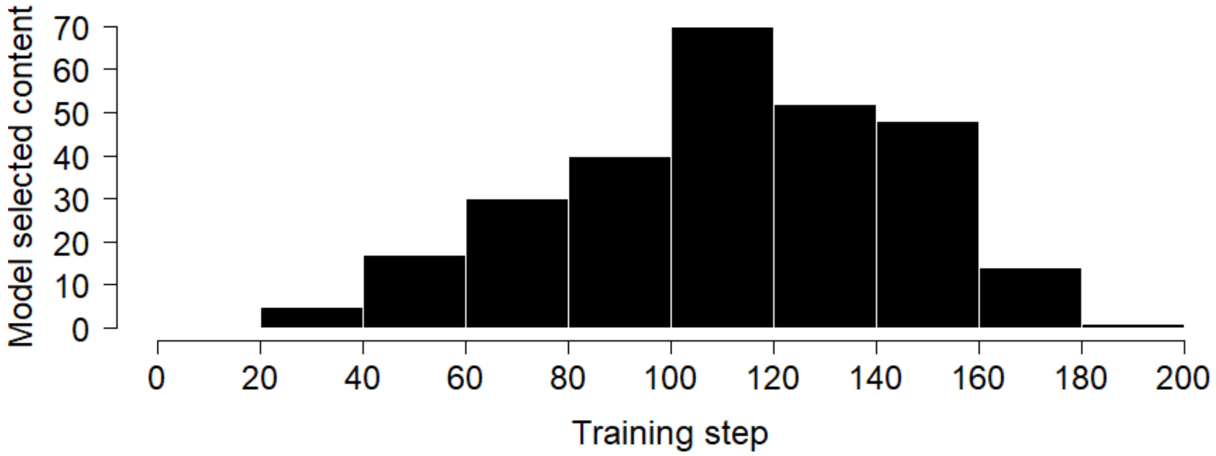
Using a robust regression (robust against violations of classic data assumptions of regression, see Maechler et al. 2020) for the training steps generating the "best" texts with highest overall quality scores on the quality score components, we observe in Table W1.5 that the content uniqueness among the top 10 ranked websites is the most important determinant for at which training step the most optimal content is generated. That means that when the top 10 ranked websites are more unique compared to each other (i.e., the top ranked websites on which we fine-tune do not make use of many common phrases) our method selects content from a later training phase (*B*=117.88, *t*=4.47, p<.000). This may arise because the risk to pick up the repetitive language patterns is lower, and additional fine-tuning steps are needed because the top search results contain more unique phrases. Interestingly, the regression model explains just ~11% of the variance in the data (Adj.R^2=.1084), meaning that the probabilistic fine-tuning and text generation processes of the GPT-2 model has a considerable impact on at which training step the most optimal content is generated.

Figure W1.5.1: Median & IQR Model Fit (Over all Trainings for Keywords)



— Median of model training loss for all model trainings; ▨ IQR of model training loss for all model trainings

Figure W1.5.2: Quantity of Model Selected Most Optimal Content vs. Training Step



■ Quantity of mean top model selected content (for each keyword, we extracted the top scoring generated content and calculated the mean training step from which these came from)

Table W1.5: Quality Score Factors Determining the Training Step for Optimal Content Selection

Robust Regression[1]				
Independent Variables	*B*	Std. Error	*t*	*p*
Intercept	54.79	26.56	2.06	.039*
Topic (s_a) + Keywords (s_k) of Top 10	4.82	6.92	0.69	.486
Uniqueness (s_d) of Top 10	117.88	26.37	4.47	<.000**
Readability similarity (s_r) + Naturality similarity (s_n) of Top 10	-31.07	9.59	-3.24	.001**
Adjusted R^2 of regression model: .1084				

[1] Dependent variable: Model training step at which most optimal content was selected based on quality score; statistical significance codes: *0.05 level, **0.01 level; because of strong pairwise correlations, we combined s_a and s_k as well as s_r and s_n into one variable by adding them up.

1.6 External Validation of Method Performance

In this section, we assess the generalizability of our proposed method across keywords and industries using our quality score measure. For this purpose, we randomly choose 338 keywords from the approximately 8,500 keywords used previously (typically 9 or 10 keywords for each of the 36 industries) and generated 338,000 pieces of content (1,000 for each single keyword), of which the method automatically selected the best scoring 338 texts (1 for each keyword). Descriptives are in Table W1.4.1, columns (3) and (4).

Table W1.6 reports the difference in medians between the machine generated content and the top 10 ranked websites for all five quality score components in bold, with Wilcoxon rank sum group comparison tests as a statistical difference indicator. We find that the raw machine outperforms the top 10 ranked content for most quality score components in all four industry sectors (Table W1.6). For example, our method outperforms the top 10 ranked websites in terms of topic consistency (s_a) by ~9% in industry sector I (+.09**), scoring at 34% in topic consistency. The uniqueness of the generated content (s_d), is the only quality indicator that shows a slightly lower value in comparison to the top 10 ranked websites (e.g., -.03** (-3%) in industry sector III), though being at a high value in absolute terms (e.g., ~87% in industry sector III).

Table W1.6: Machine vs. Top 10 Quality Score (All Industry Sectors)

Industry Sector	**Statistics**	**Topic (s_a)[1]**	**Keywords (s_k)[1]**	**Uniqueness (s_d)[1]**	**Readability similarity (s_r)[1]**	**Naturality similarity (s_n)[1]**
I.	**Raw Machine vs. Top 10[1]**	**+.09****	**+.14****	**+.03***	**+.31****	**+.25****
	Raw Machine Median[2]	.34	.34	.88	.91	.83
II.	**Raw Machine vs. Top 10[1]**	**+.08****	**+.13****	**-.02**	**+.22****	**+.24***
	Raw Machine Median[2]	.40	.40	.88	.83	.83
III.	**Raw Machine vs. Top 10[1]**	**+.10****	**+.14****	**-.03****	**+.22****	**+.07**
	Raw Machine Median[2]	.43	.44	.87	.83	.67
IV.	**Raw Machine vs. Top 10[1]**	**+.11****	**+.15****	**-.04***	**+.31****	**+.23****
	Raw Machine Median[2]	.40	.40	.88	.91	.83

[1] Difference in quality score component median value: raw machine generated content vs. real top 10 ranked websites; p-value from Wilcoxon rank sum 2-group comparison tests between machine generated content and top 10 ranked websites; statistical significance codes (one-tailed): *0.05 level, **0.01 level;
[2] Median quality score component value for raw machine generated content; n=338;

1.7 Alternative Quality Score Weighting Performance Assessment

Our quality score (qs_g) currently consists of 5 dimensions that are weighted equally as depicted in formula (1) in the main manuscript (i.e., the 5 dimensions are just multiplied to get

qs_g). To evaluate the current weighting as depicted in formula (1) in the main manuscript, we consider an alternative weighting scheme: maximize the products of s_a, s_k, s_d, while having minimum thresholds (i.e., cutoff-values) for s_n and s_r. Table W1.7 compares the scores of the current against the alternative weighting scheme for all content produced for our field experiments in the IT service industry and the education sector when applying a 50% (i.e., keep 50% of the top scoring pieces of content) and a 25% (i.e., keep 25% of the top scoring pieces of content) cutoff value for s_n and s_r. A positive (negative) value in Table W1.7 means the quality score weighting scheme presented in the article according to formula (1) performs better (worse) than the alternative quality score weighting scheme. For example, using a cutoff value of 50% in the IT service industry sector experiment, the proposed quality score weighting scheme is superior for s_a (.022**), s_k (.042**), s_d (.042**), and qs_g (.008**), which is consistent across both the used cutoff values and experimental contexts. Thus, the alternative quality score weighting we considered does not result in any improvement, as cutting off content with lower s_n and s_r often results in discarding content that performs well in terms of s_a, s_k, s_d, which ultimately translates into a lower score for the overall quality metric qs_g.

Table W1.7: Comparison of Current vs. Alternative Quality Score Weighting Scheme

Field experiment	s_n & s_r cut-off value[1]	s_a[2]	s_k[2]	s_d[2]	s_n[2]	s_r[2]	qs_g[2]
IT service	50%	.022**	.042**	.042**	-.083**	-.021**	.008**
IT service	25%	.022**	.046**	.178**	-.167**	-.043**	.022**
Education	50%	.030**	.045**	.013**	-.083**	-.043**	.003**
Education	25%	.045**	.059**	.021**	-.083**	-.021**	.014**

[1] The cut-off value specifies how many top-scoring data points to maintain, i.e., 50% means keep 50% of the top scoring data-points in s_n & s_r, 25% means keep 25% of the top scoring data points in s_n & s_r (25% is thus more conservative)

[2] Reported values are median difference values, i.e., median qs_g score of old quality score scheme minus median qs_g score of new (as suggested by the reviewer) quality score scheme. A positive value means the old scheme is superior, a negative value means the new scheme is superior. Significance codes come from two-tailed Wilcoxon rank sum 2-group comparison tests: *0.05 level, **0.01 level;

1.8 Machine Generated Sample Content

To demonstrate the versatility of our approach, Tables W1.8.1, W1.8.2, and W1.8.3 present abbreviated samples of machine-generated content for keywords from varied industries (“best e bike insurance,” “aerospace component manufacturer,” and “state library bookshop”) that have not yet been revised by a human. For comparison reasons each of Tables W1.8.1, W1.8.2, and W1.8.3 contain a real example of top 10 and worst 10 search engine ranked content (i.e., in the examples presented here we used Google’s search engine rankings 290-300). As in Table 2, the texts are shown with their associated quality score components. We can see that while lower ranked real content is both off-topic and performs poor in terms of quality score components (except for uniqueness, for reasons as already discussed in section 1.4 of this

Appendix), the machine generated content typically matches those of the top 10 samples very well and points to subtle differences not immediately transparent to human readers.

Table W1.8.1: Example Generated Piece of Content vs. Top & Worst ranking Content

Stage	Quality Score	Generated Content
GPT-2 fine-tuned (not revised)	s_a = .65 s_k = .73 s_d = .84 s_n = .58 s_r = .62 qs_g = .144	**Best e bike insurance** **Best value electric bike insurance** We offer a range of products to suit every budget and circumstance. Our products can provide additional protection for theft, damage, accident, negligence, fraud and more. Our customer service reps are here to tailor a program to your needs. **Bicycle Roadside Assistance** If you find yourself with a disabled bike well come pick you up. […]
Real top10 ranked content	s_a = .58 s_k = .61 s_d = .96 s_n = .50 s_r = .79 qs_g = .133	**The importance of e-bike insurance** Electric bikes (commonly referred to as e-bikes) are expensive, high potential risks on the road typically not fully covered by homeowners, renter or auto insurances. Its a risky move for your e-bike to be uninsured. Fortunately, our bicycle insurance is designed specifically for e-bikes to give you the peace of mind you deserve. Lets take a closer look at why e-bike insurance is a good move. E-bikes can be expensive. […]
Real worst10 ranked content	s_a = .19 s_k = .08 s_d = .96 s_n = .17 s_r = .21 qs_g = .001	**Cycle brands in mumbai** And summer seasons are the perfect time to shape yourself into an envious body, dont you think so? You would be thinking why are we talking [Continue reading] Top 20 Sportswear Brands for Men and Women in India Metro Cycle Company in Charni Road, Mumbai. Know more about latest cycles and new cycles in India 2020. [brand name] [brand name] ([Phone number]) is the leading Bicycle Dealer In India. Incense sticks nowadays are made up of aromatic materials that produce aroma when burned. Cycles in India Get all Cycles information like cycle prices, reviews, photos, etc. . There are more cycle brands in India but we have listed only 25, drop your comment for your best cycle. , one of the largest cycle manufacturers in India is now emerging as best Bicycle Companies online. […]

Human revision in our field experiment reported below: ▨ = identifiers like brand names and phone numbers are replaced by a tag (e.g., [brand name]) to retain confidentiality; headlines are printed in bold to ease reading;

Table W1.8.2: Example Generated Piece of Content vs. Top & Worst ranking Content

Stage	Quality Score	Generated Content
GPT-2 fine-tuned (not revised)	s_a = .40 s_k = .40 s_d = .88 s_n = .92 s_r = .98 qs_g = .127	**Aerospace component manufacturer** With a revenue of $8 billion, [brand name] is the third-largest aerospace company in the world. It provides products for both global aerospace and defense, manufacturing nacelle systems, avionics systems, control surfaces, electrical systems, mechanical and aerospace systems, and flight simulator systems. The company also provides aftermarket services, like draping and electrical systems, in addition to product development. [brand name] specializes in fabricating complex components and sub-assemblies for military and commercial uses. Examples of products it creates are for use in communication systems, missiles and aviation defense, and in the aerospace industry generally. […]
Real top10 ranked content	s_a = .31 s_k = .38 s_d = .99 s_n = .25 s_r = .60 qs_g = .017	**[brand name]: World Leaders in Aerospace** Since Flight Began [brand name] comprises one of the largest, most experienced concentrations of world-class aerospace companies, the world's AEROSPACE ALLEY! Our firms grew up where aerospace was born. Today, our precision components take to the skies every day, in every corner of the globe, as they have since the inception of powered flight. [brand name] member companies are united by a single goal; be the world leader in providing customers with aerospace components of unsurpassed quality, at competitive prices and always,on-time delivery. To achieve that goal, we work with our peers and with regional, national and international experts to ensure continuous improvement of our products and services. […]
Real worst10 ranked content	s_a = .07 s_k = .05 s_d = .91 s_n = .17 s_r = .13 qs_g = .000	**Aerospace Industry** **Aerospace Industry** TOTAL: 8 HORNET Series INSPIRER Series GRANDER 5Ax Series DBC2000mm DBC1500mm GRANDER 5MG Series DBC3100mm DBC2500mm Industry Aerospace Industry GRANDER 5Ax Series GRANDER 5MG Series HORNET Series INSPIRER Series DBC 1500mm DBC 2000mm DBC 2500mm DBC 3100mm Automobile HE Series HT Series VTW Series VA Series Medical Sphere VTH Series VTP Series VTJ Series VTG Series Large Molding DBC 1500mm DBC 2000mm DBC 2500mm DBC 3100mm Electronics Products HF Series VA Series VF Series VTT Series Parts Processing HE Series HF Series VTT Series VTW Series VF Series VA Series VH Series VE Series VK Series VP Series VTH Series TS Series CNC Lathes T Series CNC Lathes TEL: [phone number] FAX: [fax number] Email: [email] Address: [address] TEL: [phone number] Email: [email] Address: [address] 2018 [brand name] All rights reserved.

Human revision in our field experiment reported below: ▨ = identifiers like brand names and phone numbers are replaced by a tag (e.g., [brand name]) to retain confidentiality; headlines are printed in bold to ease reading;

Table W1.8.3: Example Generated Piece of Content vs. Top & Worst ranking Content

Stage	Quality Score	Generated Content
GPT-2 fine-tuned (not revised)	s_a = .25 s_k = .22 s_d = .88 s_n = .92 s_r = .98 qs_g = .042	**State library bookshop** open seven days a week […] $10 kids 12 and under Free playtime on the playroom terrace **Discover the wonders of Queensland** The [brand name] Historical Library is Queensland's first permanent, accessible, fully-staffed library. The Library is home to a beautifully curated selection of Queensland's best contemporary and classic books, and offers free access to more than 220,000 unique readers each day. The Library is Brisbane's only permanent non-profit library, and is funded entirely by the State Government. […]
Real top10 ranked content	s_a = .27 s_k = .25 s_d = .96 s_n = .42 s_r = .72 qs_g = .019	**Readings State Library** Readings State Library is the first permanent commercial bookshop permitted to be established in the building in the Library's 150 year history. It is a unique space that caters to students, travellers, book lovers, CBD workers and library regulars. As part of the State Library Victoria's Vision 2020 redevelopment project, our shop was relocated to a welcoming new space beautifully designed by Australasian design firm Architectus with Scandinavia's Schmidt Hammer Lassen Architects. […]
Real worst10 ranked content	s_a = .11 s_k = .03 s_d = .96 s_n = .25 s_r = .74 qs_g = .000	**About the Author** Tasmanian writer Katherine Johnson is the author of four novels. Paris Savages based on a true story of human zoos will be published in October 2019. Katherines third novel, Matryoshka or Russian dolls (Ventura Press 2018), is a story of secrets, refuge and loves lost and found. Her previous novels include The Better Son (Ventura Press 2016), set in northern Tasmanias caves, and Pescadors Wake (Fourth Estate 2009), the story of the danger and heartbreak of lives at the mercy of the sea during a three-week Southern Ocean chase. Katherine is the recipient of The University of Tasmania Prize and the Peoples Choice Award (Tasmanian Literary Prizes) and HarperCollins Varuna Awards. The Better Son was longlisted for the Australian Indie Book Awards 2017 and the Tasmania Book Prize (Premiers Literary Prizes 2017). […]

Human revision in our field experiment reported below: ▨ = identifiers like brand names and phone numbers are replaced by a tag (e.g., [brand name]) to retain confidentiality; headlines are printed in bold to ease reading;

2 Appendix References

Baayen RH, Shafaei-Bajestan E (2019) Analyzing linguistic data: A practical introduction to statistics. Package ‘languageR’. Version 1.5.0. *CRAN*. Accessed May 20, 2019, https://cran.r-project.org/web/packages/languageR/languageR.pdf

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A, (2018) “quanteda: An R package for the quantitative analysis of textual data.” *Journal of Open Source Software*. 3(30). https://doi.org/10.21105/joss.00774

Berger J, Sherman G, Ungar L (2020b) TextAnalyzer. Accessed November 11, 2020, http://textanalyzer.org

Bronnenberg BJ, Kim JB, Mela CF (2016) Zooming in on choice: How do consumers search for cameras online? *Marketing Science*. 35(5):693-712.

Danaher PJ, Mullarkey GW, Essegaier S (2006) Factors affecting website visit duration: A cross-domain analysis. *Journal of Marketing Research*. 43(2):182-194.

Edelman B, Zhenyu L (2016) Design of search engine services: Channel interdependence in search engine results. *Journal of Marketing Research*. 53(6):881-900.

Flanigan, AJ, Metzger, MJ (2007) The role of site features, user attribtues, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*. 9(2):319-342. https://doi.org/10.1177/1461444807075015

Jerath K, Ma L, Park YH (2014) Consumer click behavior at a search engine: The role of keyword popularity. *Journal of Marketing Research*. 51(4):480-486.

Kamoen N, Holleman B, Bergh H (2013) Positive, negative, and bipolar questions: The effect of question polarity on ratings of text readability. *Survey Research Methods*. 7(3):181-189.

Liu J, Toubia O (2018) A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*. 37(6):930-952.

Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Conceicao ELT, Palma MA (2020) Basic robust statistics. Package ‘robustbase’. Version 0.93-6. *CRAN*. Accessed May 20, 2020, https://cran.r-project.org/web/packages/robustbase/robustbase.pdf

Pennebaker JW, Booth RJ, Boyd RL, Francis ME (2015) Linguistic inquiry and word count: LIWC2015. Austin, TX: Pennebaker Conglomerates. Accessed November 1, 2020, www.LIWC.net.

Pitler E, Nenkova A (2008) Revisiting Readability: A unified framework for predicting text quality. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 186-195.

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. OpenAI.

Roberts C (2010) Correlations among variables in message and messenger credibility scales. *American Behavioral Scientist*. 54(1):43-56.

Rocklage MD, Rucker DD, Nordgren LF (2018) Persuasion, emotion and language: the intent to persuade transforms language via emotionality. *Psychological Science*. 29(5):749-760.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomze AN, Kaiser L, Polosukhin I (2017) Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. 1-15.