Supporting Content Marketing with Natural Language Generation

WEB APPENDIX

Index

1 Technical Modeling and Validation Notes 4

1.1 GPT-2 Model Description 4

1.2 Description of Essential Software Features 6

1.3 Applied Uniqueness, Naturality & Readability Measures 8

1.4 External Validation of Method Assumptions and Quality Score 9

1.5 Validation of Method Fine-Tuning Process 14

1.6 External Validation of Method Performance 17

1.7 Alternative Quality Score Weighting Performance Assessment................................... 18

1.8 Machine Generated Sample Content 20

2 IT Service Sector Field Study 23

2.1 Keywords to Optimize for & Keyword Stats 23

2.2 Experimental Setup 24

2.2.1 Experimental Quality Assurance 24

2.2.2 Stimuli for Content Production 26

2.2.3 Content Production: Text Lengths and Human Revision Change 29

2.2.4 A/B Testing Experiment for Content Improvement 30

2.3 Supplemental Content Performance Tests 33

2.3.1 Quality Scores of Experimental Groups' Content 33

2.3.2 Post Hoc Test for Achieved Search Engine Rankings 35

2.3.3 Content Ranking for Sub-Keywords Assessment 36

2.3.4 Additional Keywords Performance 38

2.4 Consumer Content Perceptions: MTurk Study 41

2.4.1 Survey Instructions 41

2.4.2 Operationalization, Covariate Checks & Data Quality Assurance 42

2.4.3 Content Perception Inter-Correlations 44

2.5 Consumer Content Perceptions: Computational Analyses 45

2.6 Website Engagement 47

2.7 Content Production Costs Calculation Details 49

3 Education Sector Field Study 51

3.1 Keywords to Optimize for & Keyword Stats 51

3.2 Supplemental Content Performance Tests 53

4 Keyword Boundary Conditions 55

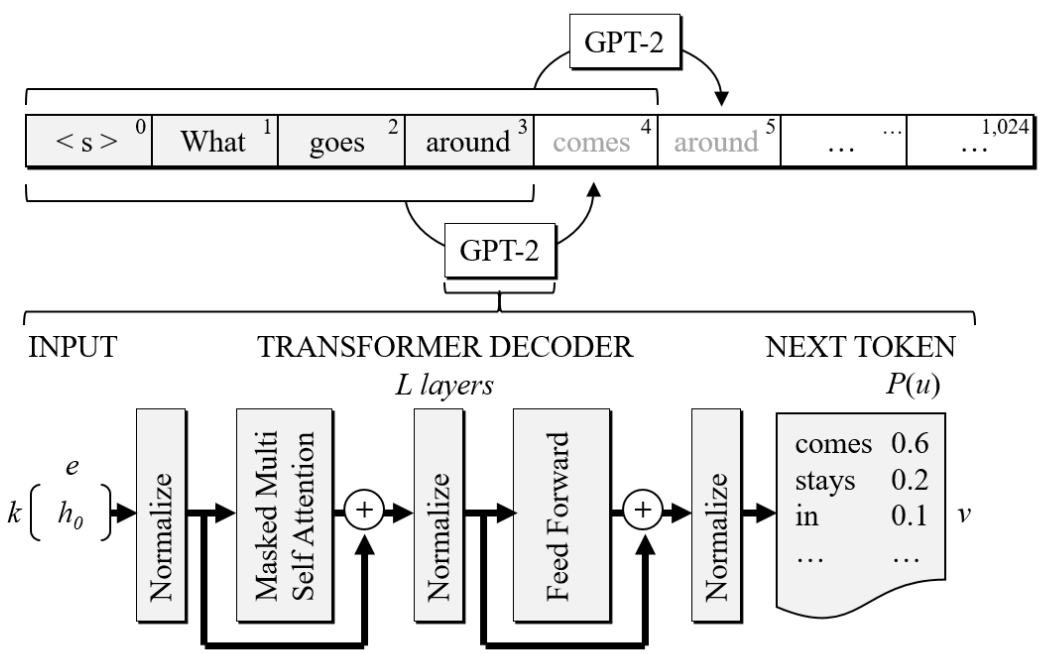
5 Appendix References 56

1 Technical Modeling and Validation Notes

1.1 GPT-2 Model Description

To provide sense of 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, $U=(u_{-k},\ldots,u_{-1})$, 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 pieces of words).

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

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

$$h_0 = UW_e + W_p \quad \text{(W2)}$$

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

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

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, 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 given sequence words 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, e.g., 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), 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) 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 section 2.2. of the paper could be adapted to incorporate other linguistic components. In 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.

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

Content that fails to achieve this level of uniqueness is discarded. 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 a trial and error 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~.82 (i.e., 82% unique), and a n-gram-size of 7~.88 (i.e., 88% unique).

Naturality similarity measures (s_n). For quantifying the naturality similarity metric, we applied 12 linguistic measures which assess the lexical richness and composition of a text using the "languageR" package developed by Baayen and Shafaei-Bajestan (2019). 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, calculation 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 "quanteda" package developed by Benoit et al. (2020). 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. (2020).

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 W1.4a, columns (1) and (2). Using Wilcoxon rank sum group comparison tests, Table W1.4b 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 W1.4b 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. This may arise because the top 10 ranked websites consistently reflect similar topics as opposed to lower ranked websites. Thus, we can ascertain that fine-tuning on the top 10 ranked websites' content will produce the most optimal content, and approve 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 W1.4a: 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 W1.4b: 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 W1.4b 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 W1.4c) and for specific industries (e.g., tourism) (Table W1.4d). 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 W1.4c: External Validation of Method Assumptions Statistics for Industry Sector III as Example

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;

Table W1.4d: External Validation of Method Assumptions for the Tourism Sector as Example

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;

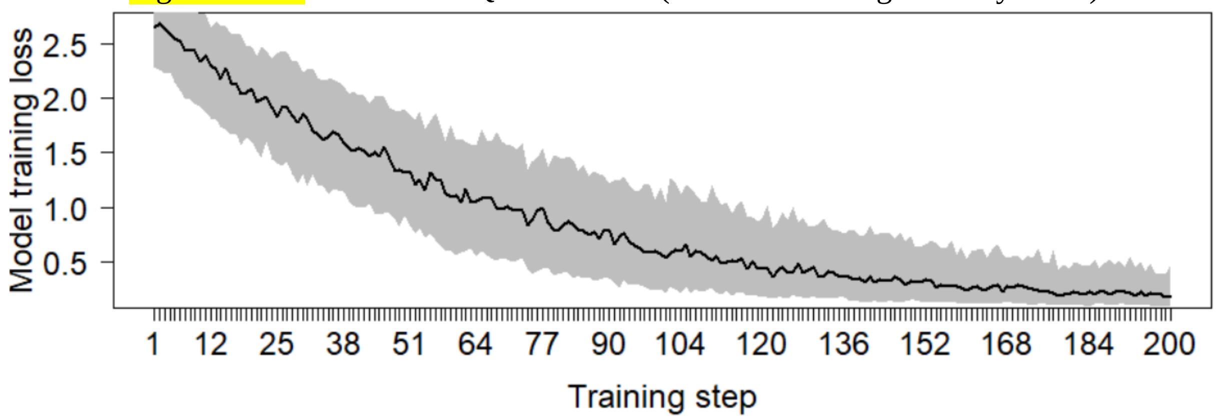
1.5 Validation of Method Fine-Tuning Process

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 above proposed quality score. 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 W1.5a 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 main text. While model fit is consistently improving, Figure W1.5b 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.

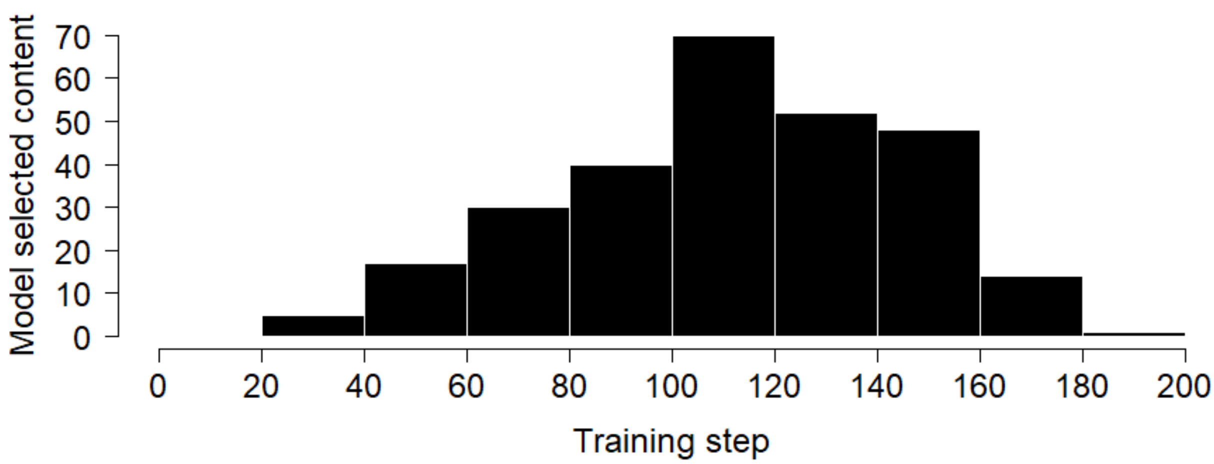
Using a robust regression (robust against violations of classic data assumptions of regression, see Maechler et al. 2020) for the training steps generating the "best" texts with highest overall quality scores on the quality score components, we observe in Table W1.5 that the content uniqueness among 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 W1.5a: 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 W1.5b: 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 W1.4a, 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 1.7 compares the scores of the current vs. the alternative weighting scheme for all content produced for our field experiments in the IT service and education sectors 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) percent cutoff value for s_n and s_r. A positive (negative) value in Table 1.7 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, for 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 for both 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.

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 sn & sr, 25% means keep 25% of the top scoring data points in sn & sr (25% is thus more conservative)
[2]Reported values are median difference values, i.e., median qs score of old quality score scheme minus median qs 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 W1.8a, W1.8b, 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 W1.8a, W1.8b, contain a real example of top 10 and worst 10 search engine ranked content (i.e., in the examples used here we used Google’s search engine rankings 290-300). As in Table 2, the texts are presented 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, 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 W1.8a: 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 W1.8b: 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;

2 IT Service Sector Field Study

2.1 Keywords to Optimize for & Keyword Stats

Table W2.1 depicts the keywords (search queries) the experimental groups in the IT service sector field study produced optimal content for, including keyword statistics (e.g., the average monthly search volume, the paid keyword competition, the keyword length), and descriptive statistics for the ranking performance of the revised machine content in the search engine. The company selected the keywords used in our experiment based on its standard procedure (i.e., based on monthly search volume, competition, fit with the firm and keyword-strategy). The company tended to select longer tailed, lower search volume keywords since that fits their keyword strategy.

Table W2.1: Keywords for IT Service Field Experiment

Keyword	Avg. monthly search volume	Competition	Competition index	Keyword length	Mean revised machine ranking	Median revised machine ranking	SD revised machine ranking	IQR revised machine ranking	% of days revised machine was in ranking
	Descriptives								
IT procurement	10	low	4	2	11.65	8	7.68	10	90.70
IT support and services	10	low	3	4	15.61	15	6.54	5	97.21
global IT support	10	-	-	3	15.03	14	5.11	3	72.09
IT assessment	10	low	0	2	21.58	22	5.87	6	87.91
IT consulting services	10	-	-	3	18.22	17	7.43	5	66.05
IT maintenance	10	-	-	2	13.66	12	8.11	11	99.07
IT service maintenance	0	-	-	3	3.19	2	3.25	2	99.07
IT service support	10	low	0	3	10.74	10	2.82	4	99.07
IT service continuity	10	-	-	3	21.57	19	6.47	9	92.56
IT support business	10	-	-	3	51.60	52	14.29	26	39.53
Small business IT support services	0	-	-	5	94.75	50	79.96	129	64.19
IT support costs for small business	0	-	-	6	14.21	13	3.96	6	99.07
IT maintenance support	0	-	-	3	4.11	2	3.86	6	99.07
IT maturity assessment	10	low	29	3	22.67	22.5	7.10	9.75	97.67
IT procurement services	10	-	-	3	3.23	3	1.11	2	99.07
IT procurement process	10	-	-	3	30.18	19	22.22	21	99.07
IT solution delivery	10	low	0	3	2.47	1	2.21	2	90.70
IT strategy consulting	10	-	-	3	28.25	26	9.24	5	98.14
IT consulting software	10	high	100	3	15.70	15	5.16	8	75.35

Entries that display “-“ mean that the search engine keyword tool did not provide specific information.

2.2 Experimental Setup

2.2.1 Experimental Quality Assurance

Four experimental groups produced content for the company’s website. The groups consist of (1) 19 novices (untrained marketing students who received a written stimulus that broadly stated the task), (2) 19 quasi-experts (marketing students who were trained in class and received a written instruction and a clear direction of how to do it), (3) 5 SEO experts (professionals with at least two months experience in the SEO industry who received the novices' stimulus[3]), and (4) the semi-automated SEO content writing machine with revisions made by a company employee who was instructed to keep content changes to a minimum.

Groups (1) – (3) produced content via an online survey that contained a link to the keyword-specific top 10 search engine ranked content and a word-counter tool that the company uses in its usual content production workflow. Thus, we gave participants all tools and the environment that the company uses in its common SEO content production workflow.

Participants in the experiment were given incentives for content production. The incentive for Groups 1 (Novices) and 2 (Quasi-experts) was 15 € per produced content and credit for a marketing course. The incentive for Group 3 (SEO experts) was 40 € per produced content.

The selected keywords are uniform in terms of keyword statistics (competition, search volume, keyword length), and the company did not have any prior history of ranking for these in the search engine. In addition, we checked if the real SEO experts were able to improve their initial content draft by providing them the opportunity to improve their initial draft using the quality score information of their and of the top10 ranked content, which was not the case (Section 2.2.4).

Content production took place within the same week and in the same geographic location so that all participants had the same state of search engine results as a basis, which we controlled for via daily crawls.

To control for content length across the groups, we provided participants with a guideline on text length in terms of number of words. Based on a Kruskal Wallis group comparison, the

[3] Survey instructions are reported in Table W6 in the Web Appendix B1.

human content writing groups did not differ in their education ($\chi^2(3)=.60$, $\eta^2=.01$, $p=.745$) or writing skills (for which the SEO experts scored a bit higher; $\chi^2(3)=5.89$, $\eta^2=.12$, $p=.053$), and the time invested conducting research on the target keyword / topic ($\chi^2(3)=.28$, $\eta^2=.00$, $p=.868$) and writing ($\chi^2(3)=3.76$, $\eta^2=.08$, $p=.153$). Descriptive statistics on the content length and changes are provided in Table W2.2.3 (Section 2.2.3). These show that the produced content between the experimental groups is of equal length and that in the revision process, the human reviser changed ~9% of the machine-made content. In addition, Table W2.3.1 (Section 2.3.1) illustrates that the machine-made content decreased in terms of search engine optimization (quality score measured) due to the human revision, though the revised machine still performs better than the human content producing groups in terms of the quality score.

2.2.2 Stimuli for Content Production

Table W2.2.2 reports the stimulus for the content writing groups in our IT service industry experiment.

Table W2.2.2: Participants' Survey Instructions for Content Writing

Content Writing Group	**Instructions**[1]
Novices	[Short introduction stating the goal of this study, strict anonymization, the incentive and a contact person for questions.] Imagine you are a marketing employee in an IT service company. Your manager approaches you to **write a Google search engine optimized (SEO) text for a single site on the website of your IT company**, that elaborates on a specific service. You should **write the text in a way that it ranks well in Google**. That means, it should preferably appear on page 1 in the Google search results. • The text should be written for the keyword / search term / topic: "**IT maintenance**" (i.e., for IT maintenance provided as a service by your company to firms). • It should be written for ranking well in **Google in [Country blinded], set to English language** (please use the link below). • For ranked **example sites see**: https://www.google.com/search?num=100&hl=en&q=it+maintenance • It should be **original, unique content**, invented by you (i.e., NO copies). • It should be written **in English language**. • It should contain **around 700 to 800 words** (ca. 2 A4 pages). **Your text:** (Please write your text in the following text field.)
Quasi Experts	[Short introduction stating the goal of this study, strict anonymization, the incentive and a contact person for questions.] Imagine you are a marketing employee in an IT service company. Your manager approaches you to **write a Google search engine optimized (SEO) text for a single site on the website of your IT company**, that elaborates on a specific service. You should **write the text in a way that it ranks well in Google**. That means, it should preferably appear on page 1 in the Google search results. • The text should be written for the keyword / search term / topic: "**IT maintenance**" (i.e., for IT maintenance provided as a service by your company to firms). • It should be written for ranking well in **Google in [Country blinded], set to English language** (please use the link below).

- For ranked **example sites see**: https://www.google.com/search?num=100&hl=en&q=it+maintenance
- It should be **original, unique content**, invented by you (i.e., NO copies).
- It should be written **in English language**.
- It should contain **around 700 to 800 words** (ca. 2 A4 pages).

How to write a SEO optimized text?

- **Integrate the main keyword** ("IT maintenance") **or parts of it most often compared to the other words in your text**.
- **Write about subtopics / content** that you can find on the top ranked websites for the main keyword.
- **Align the word distribution of your text** with the word distribution of the top ranked websites for the main keyword (i.e., **put the right words with the right frequencies into your text**).
- For the **word distribution analyses use**: https://wordcounter.net/ (Please be aware that the tool doesn't count common stopwords like "it".)
- **Prevent keyword stuffing** (i.e., don't integrate keywords overly often and in an unnatural way into your text).
- Try to give your text a **good readability and structure**.

Your text: (Please write your text in the following text field.)

Real SEO Experts	[Short introduction stating the goal of this study, strict anonymization, the incentive and a contact person for questions.]

Imagine you are a marketing employee in an IT service company.

Your manager approaches you to **write a Google search engine optimized (SEO) text for a single site on the website of your IT company**, that elaborates on a specific service. You should **write the text in a way that it ranks well in Google**. That means, it should preferably appear on page 1 in the Google search results.

- The text should be written for the keyword / search term / topic: "**IT maintenance**" (i.e., for IT maintenance provided as a service by your company to firms).
- It should be written for ranking well in **Google in [Country blinded], set to English language** (please use the link below).
- For ranked **example sites see**: https://www.google.com/search?num=100&hl=en&q=it+maintenance
- It should be **original, unique content**, invented by you (i.e., NO copies).

- It should be written **in English language**.
- It should contain **around 700 to 800 words** (ca. 2 A4 pages).

Your text: (Please write your text in the following text field.)

[1]Keywords and links were adapted in each survey.

2.2.3 Content Production: Text Lengths and Human Revision Change

To control for content length across the groups, we provided participants in our experiment with a guideline on text length in terms of a target length (number of words) for content production. Table W2.2.3 shows descriptive statistics for the length of produced content (as number of words), illustrating that the produced content between the experimental groups is of equal length, hovering around 700 to 800 words. In addition, Table W.2.2.3 illustrates that the human reviser that corrected the machine generated content usually corrected around 9.04% (~74 words) of the content.

Table W2.2.3: Descriptives for Content Lengths and Revision

Dimension	**Groups**	**Descriptives**			
		Median	(IQR)	Min	Max
Produced content length (in words)	Revised machine	807	(67)	632	899
	Real SEO Experts	729	(84)	578	771
	Quasi Experts	694	(69.5)	498	749
	Novices	711	(48.5)	377	966
Content change (raw vs. revised)[1]	Change in %	9.04	(3.77)	3.31	21.45
	Change in words	74.00	(36.50)	27.00	154.00

[1]This includes every possible change between the raw machine and revised machine output like added words, deleted words, and words with at least one changed letter (including changed letter capitalization).

2.2.4 A/B Testing Experiment for Content Improvement

We conducted an additional study to provide the real SEO experts with an opportunity to improve their previously produced content by being provided with the quality scores of their initial content and the real search engine ranked top 10 content. The experimental setup is as follows: The A/B tests were done using an online survey (including personal explanations of the task and a Q&A by the researcher). The online survey contained a broad information on the task and researcher contact details, and an incentive of 40€ for the revision / feedback round per piece of content. In addition to being provided with a task description and an introduction to the quality score and its components, they were also offered an explanation of each component of the quality score and how to interpret the reported quality scores numbers. They also received feedback on the original content that they had produced and the quality score it achieved, both on each component and overall. For comparison purposes, they were also provided with the top 10 ranked content for the specific keyword and their associated quality scores. The revised text was entered in an open text field. We extended the original study reported in the main manuscript using 30 keywords for the real SEO experts (instead of just 9), so the A/B testing was done for 30 pieces of real SEO expert produced content.

Table 2.2.4a shows that the real SEO experts changed their original content by 10.24% (~77.50 words), ranging between 12 words changes and 176 word changes. Table 2.2.4b compares the achieved quality scores of the original SEO experts' content to the revised SEO experts' content for each quality score component. We find no statistically significant differences between them, suggesting that the SEO experts were not able to improve the quality of their content, likely due to the associated complexity (i.e., dozens of word distributions, numbers, and abstract concepts). This suggests that the semi-automated procedure not only reduces the time/cost associated with content production, but also performs better than human experts on tasks involving the generation of content for a specific purpose.

Table 2.2.4a: Descriptives for Real SEO Experts Content Revision

Dimension	Groups	Descriptives			
		Median	(IQR)	Min	Max
Produced content length (in words)	Original Real SEO Experts	729.5	(47.25)	587	819
	Revised Real SEO Experts	760.5	(58.75)	546	930
Content change (original vs. revised)[1]	Change in %	10.24	(5.84)	1.62	24.24
	Change in words	77.50	(49.25)	12	176

[1]This includes every possible change between the original real SEO experts and revised real SEO experts content like added words, deleted words, and words with at least one changed letter (including changed letter capitalization).

Table 2.2.4b: Quality Score: Original Real SEO Experts Content vs. Revised Real SEO Experts Content

Quality Score Component	Group	Descriptives			Wilcoxon rank sum[1]			
		Median (IQR)	Min	Max	W	z	*r*	*p*
Topic (s_a)	Original Real SEO Experts	**.41 (.18)**	**.30**	**.72**				
	Revised Real SEO Experts	.44 (.18)	.30	.72	423	.39	.05	.697
Keywords (s_k)	Original Real SEO Experts	**.46 (.20)**	**.16**	**.79**				
	Revised Real SEO Experts	.46 (.15)	.16	.80	423	.39	.05	.697
Uniqueness (s_d)	Original Real SEO Experts	**.97 (.06)**	**.08**	**1.00**				
	Revised Real SEO Experts	.96 (.07)	.63	.99	560	-1.62	-.21	.105
Readability similarity (s_r)	Original Real SEO Experts	**.72 (.54)**	**.02**	**1.00**				
	Revised Real SEO Experts	.66 (.71)	.02	1.00	469	-.27	-.04	.784
Naturality similarity (s_n)	Original Real SEO Experts	**.67 (.39)**	**.08**	**1.00**				
	Revised Real SEO Experts	.58 (.25)	.17	1.00	477.5	-.40	-.05	.687

[1]Two-tailed tests between original vs revised real SEO experts quality scores, statistical significance codes: *0.05 level, **0.01 level;

2.3 Supplemental Content Performance Tests

2.3.1 Quality Scores of Experimental Groups' Content

In Table W2.3.1, we compare the quality score components from each experimental group and the top 10 ranking search results. The topic (s_a), keyword (s_k), and readability similarity (s_r) scores are higher for the raw and semi-automated content compared to the remaining experimental groups and the top 10 ranked websites as well as the lowest ranked search results, while human created content scores higher in uniqueness. That indicates that the machine may reflect or mimic patterns such as contained topics, keywords and readability levels in the top10 more thoroughly.

Table W2.3.1: Quality Score Components Group Comparisons to Top 10 Ranked Websites

Quality Score Component	Group	Descriptives				Wilcoxon rank sum[1]			
		Median	(IQR)	Min	Max	W	z	r	p
Topic (s_a)	**Top 10**	**.32**	**(.11)**	**.11**	**.27**				
	Revised machine	.40	(.13)	.35	.68	65	3.36	.54	.000**
	Raw Machine	.46	(.13)	.33	.61	62	3.60	.58	.000**
	Real SEO Experts	.37	(.08)	.30	.49	59	1.27	.24	.205
	Quasi Experts	.36	(.10)	.10	.61	139	1.19	.19	.234
	Novices	.29	(.08)	.20	.56	211	-.87	-.14	.385
	Worst 10	.19	(.07)	.11	.28	344	-5.39	-.87	.000**
Keywords (s_k)	**Top 10**	**.30**	**(.16)**	**.05**	**.26**				
	Revised machine	.44	(.18)	.32	.74	68	3.27	.53	.001**
	Raw Machine	.48	(.18)	.31	.62	64	3.53	.57	.000**
	Real SEO Experts	.36	(.06)	.16	.51	67	.87	.16	.383
	Quasi Experts	.38	(.13)	.01	.70	142	1.10	.17	.271
	Novices	.31	(.22)	.52	.61	213	-.93	-.15	.354
	Worst 10	.16	(.11)	.05	.26	335	-4.97	-.81	.000**
Uniqueness (s_d)	**Top 10**	**.92**	**(.04)**	**.79**	**.97**				
	Revised Machine	.90	(.06)	.81	1.00	216	-1.02	-.16	.307
	Raw Machine	.84	(.12)	.52	.94	301	-3.66	-.59	.000**
	Real SEO Experts	.98	(.03)	.93	1.00	15	3.45	.65	.000**
	Quasi Experts	.99	(.03)	.86	1.00	58	3.56	.58	.000**
	Novices	.98	(.07)	.79	1.00	86	2.74	.45	.006**
	Worst 10	.95	(.04)	.89	.98	108	2.11	.34	.034*
Readability Similarity (s_r)	**Top 10**	**.56**	**(.08)**	**.50**	**.74**				
	Revised Machine	.87	(.17)	.47	1.00	21	4.64	.75	.000**
	Raw Machine	.96	(.09)	.70	1.00	2	5.21	.84	.000**
	Real SEO Experts	.57	(.51)	.21	1.00	77	.39	.07	.694
	Quasi Experts	.53	(.39)	.10	1.00	206	-.73	-.12	.465
	Novices	.57	(.42)	.08	.96	176	-.11	-.02	.907
	Worst 10	.47	(.10)	.26	.68	312.5	-3.84	-.62	.000**
Naturality Similarity (s_n)	**Top 10**	**.56**	**(.05)**	**.50**	**.63**				
	Revised Machine	.67	(.38)	.17	1.00	145	1.02	.16	.306
	Raw Machine	.92	(.38)	.42	1.00	55.5	3.66	.59	.000**
	Real SEO Experts	.75	(.33)	.33	1.00	25	2.96	.56	.003**
	Quasi Experts	.75	(.38)	.17	.83	132.5	1.39	.23	.164
	Novices	.58	(.38)	.00	1.00	162	.53	.09	.598
	Worst 10	.35	(.15)	.18	.53	359	-5.20	-.84	.000**

[1]Two-tailed tests; statistical significance codes: *0.05 level, **0.01 level;

2.3.2 Post Hoc Test for Achieved Search Engine Rankings

Table W2.3.2 reports results of a Kruskal Nemenyi post hoc test to compare the differences between each pair of experimental groups in search engine ranking performance as reported in Table 2 in the main manuscript. We find that the search engine performances of all experimental groups are statistically different at the 0.05 level (Table W2.3.2).

Table W2.3.2: Post Hoc Tests: Search Engine Rankings Performance Comparison (IT Service Sector)

Dimension	**Group**	**Kruskal Nemenyi Post Hoc Test (*p*)**		
		Real SEO Experts	Quasi Experts	Novices
Pages in ranking / day	Revised Machine	<.000**	<.000**	<.000**
	Real SEO Experts		.014*	<.000**
	Quasi Experts			<.000**
Pages in top 10 / day	Revised Machine	<.000**	<.000**	<.000**
	Real SEO Experts		.003**	<.000**
	Quasi Experts			<.000**
Mean rankings / day	Revised Machine	<.000**	<.000**	<.000**
	Real SEO Experts		<.000**	<.000**
	Quasi Experts			<.000**

[1]Statistical significance codes: *0.05 level, **0.01 level, chi-square approximated;

2.3.3 Content Ranking for Sub-Keywords Assessment

In search engine optimization (SEO), content is usually optimized for a single main keyword (search query). However, in search engine advertising (SEA) ads and bids are optimized for multiple keywords at once. Thus, a company could benefit from optimizing SEO content for multiple keywords at once. That is why we conducted an additional experiment for our IT service field to find out how well the experimental groups' SEO content ranks for related keywords. We identified related keywords by analyzing the word distributions of the top 10 search engine ranked content for the 19 main keywords specified in Table W2.1 and extracted the most frequent keywords and groups of words, yielding 207 related keywords. For example, when analyzing the top 10 search results for the keyword "IT assessment", we find the (most frequently occurring) following related keywords in their word distributions: "IT assessments", "business continuity", disaster recovery", "security assessment", "assessment services", "IT assessment services", "risk security assessment", "information technology assessment", "disaster recovery plan". After scraping the search engine rankings for these 207 related keywords, we find that the revised content machine ranks substantially better and more often for related keywords than the competing human groups (Table W2.2.3). For example, the revised machine ranked for 34 related keywords, and occurred in the top 10 results six times. The median search engine ranking of the revised machine is 23.

The method seems to perform surprisingly well for related keywords. In contrast to relying on heuristics such as keyword density, the fine-tuning process of our semi-automated algorithm appears to capture the overall topic and related sub-topics within the content. Thus, in the process of generating content for a specific keyword, our content also performs reasonably well in terms of search engine rankings for topic-related sub-keywords for which it was not primarily optimized. We leave the generalization of our approach to optimize for multiple keywords simultaneously as a topic for future work. This could be achieved by fine-tuning on top ranking content of many related keywords instead of fine-tuning on top ranked content for just one main keyword, and modifying the quality score function to differentiate between the main keyword and sub-keywords.

Table W2.3.3: Ranking performance of Content for Related Keywords

Group	Descriptives			
	Median (IQR) search engine ranking		Total number of ranked pages	Number of pages ranked in top10
Revised machine	**23**	**(23.00)**	**34**	**6**
Real SEO Experts	26	(35.50)	4	1
Quasi Experts	188.5	(78.25)	4	0
Novices	68	(27.50)	3	0

Desriptives for achieved rankings per experimental group for topic-related sub-keywords extracted from the top10 ranked pages (207 sub-keywords, and 51,995 total ranked pages).

2.3.4 Additional Keywords Performance

In this section, we assess the content machine's performance for additional keywords in the IT service sector experiment, which were not included in our original study, bringing the keyword count for all experimental groups to 30 keywords. Due to technical issues with the company's website that were beyond our control, we were unable to put the generated content online to evaluate search engine rankings. Nonetheless, we report the quality scores for the machine and human made content.

Table W2.3.4a depicts the additional keywords for which content was generated. Table W2.3.4b illustrates that the raw and revised machine content substantially outperforms all competing human content producing groups including the SEO experts, similar to the results depicted in Table W2.3.1.

Table W2.3.4a: Additional Keywords for the IT Service Field Experiment

Field Study	Keyword	Descriptives			
		Avg. monthly search volume	Competition	Competition index	Keyword length
IT service	SLA contract	10	low	0	2
	SLA ITIL	10	low	14	2
	service level agreement best practices	10	low	0	5
	IT security services	10	-	-	3
	ITIL incident	20	low	11	2
	ITIL ITSM	10	low	32	2
	IT maintenance contract	10	-	-	3
	IT project management	40	low	18	3
	IT scalability	10	low	0	2
	IT performance management	10	low	26	3
	Server maintenance	20	low	0	2

Entries that display "-" mean that the search engine keyword tool did not provide specific information.

Table W2.3.4b: Quality Score Components Group Comparisons to Top 10 Ranked Websites (Keyword Count Increased to 30 Keywords)

Quality Score Component	Group	Descriptives				Wilcoxon rank sum[1]			
		Median	(IQR)	Min	Max	W	z	*r*	*p*
Topic (s_a)	**Top 10**	**.38**	**(.23)**	**.21**	**.69**				
	Revised machine	.49	(.22)	.35	.71	253	2.91	.38	.004**
	Raw Machine	.50	(.19)	.28	.77	260	2.84	.37	.004**
	Real SEO Experts	.40	(.19)	.30	.72	302	2.01	.26	.043*
	Quasi Experts	.41	(.19)	.10	.68	381	.81	.10	.420
	Novices	.33	(.16)	.17	.64	490	-1.08	-.14	.281
	Worst 10	.18	(.07)	.09	.28	870	-7.23	-.93	.000**
Keywords (s_k)	**Top 10**	**.38**	**(.26)**	**.14**	**.77**				
	Revised machine	.52	(.28)	.32	.81	250	2.95	.38	.003**
	Raw Machine	.55	(.23)	.29	.85	260	2.84	.37	.004**
	Real SEO Experts	.46	(.21)	.16	.79	318	1.77	.23	.077
	Quasi Experts	.43	(.21)	.01	.77	382	.79	.10	.429
	Novices	.33	(.25)	.05	.72	498	-1.20	-.16	.230
	Worst 10	.15	(.10)	.03	.26	853	-6.77	-.87	.000**
Uniqueness (s_d)	**Top 10**	**.92**	**(.09)**	**.72**	**.99**				
	Revised Machine	.91	(.13)	.74	1.00	418	.47	.09	.641
	Raw Machine	.81	(.12)	.52	.97	694	-3.70	-.48	.000**
	Real SEO Experts	.97	(.06)	.68	1.00	203	3.51	.46	.000**
	Quasi Experts	.99	(.03)	.87	1.00	101	5.07	.66	.000**
	Novices	.99	(.04)	.79	1.00	161	4.02	.53	.000**
	Worst 10	.94	(.05)	.85	.99	323	1.87	.24	.061
Readability similarity (s_r)	**Top 10**	**.56**	**(.08)**	**.48**	**.74**				
	Revised Machine	.82	(.19)	.26	1.00	104	5.11	.66	.000**
	Raw Machine	.95	(.14)	.70	1.00	6	6.57	.85	.000**
	Real SEO Experts	.70	(.53)	.02	1.00	343	1.39	.18	.165
	Quasi Experts	.49	(.43)	.04	1.00	520	-1.28	-.17	.200
	Novices	.53	(.45)	.02	.96	497	-1.19	-.16	.234
	Worst 10	.43	(.17)	.12	.68	827.5	-5.57	-.72	.000**
Naturality similarity (s_n)	**Top 10**	**.56**	**(.06)**	**.50**	**.65**				
	Revised Machine	.67	(.40)	.17	1.00	384	.97	.13	.332
	Raw Machine	.75	(.42)	.33	1.00	227.5	3.29	.43	.000**
	Real SEO Experts	.67	(.33)	.08	1.00	346.5	1.34	.17	.181
	Quasi Experts	.50	(.33)	.17	.83	448.5	-.20	-.03	.843
	Novices	.54	(.44)	.00	1.00	453.5	-.51	-.07	.607
	Worst 10	.30	(.15)	.08	.53	896	-6.59	-.85	.000**

[1]Two-tailed tests; statistical significance codes: *0.05 level, **0.01 level;

2.4 Consumer Content Perceptions: MTurk Study

2.4.1 Survey Instructions

Table W2.4.1 reports the instructions for the survey participants.

Table W2.4.1: Participants' Survey Instructions

Survey Instructions
Dear study participant Thank you for participating in our study on SEO & text writing. Your input is vital for us. In the following, besides answering some demographic questions, **we will ask you to read and assess 1 text**. **It will take you 5 minutes at most to finish the survey.** Please **read all questions and the text mindfully and completely**, and **answer all questions as honestly and spontaneously as possible**. Follow your intuition, there are **no right or wrong answers**. All information that you provide to us will be **strictly treated as anonymous**. Thank you for your kind support. Sincerely, […] [New survey page] **Imagine, you are looking for an IT service for your company, and you come across a website with the text below. Please take a look at it.** [Randomized piece of content] [Questions to assess content]

2.4.2 Operationalization, Covariate Checks & Data Quality Assurance

To assure data quality in our survey-based content consumer perception experiment, we implemented honeypots (for antispam), attention and honesty checks (i.e., reverse coded items and same questions worded a bit differently), and excluded all surveys with a completion time lower than 1.50 minutes, leaving us with 551 surveys for our analyses. We performed scale reliability checks using Cronbach's Alpha including deleting offset items. Using a series of Kruskal Wallis tests, we assured that participants' properties did not differ substantially between the experimental conditions in terms of the time to finish the survey ($\chi^2(3)=3.38$, $\eta^2=.01$, $p=.337$), the participants' gender ($\chi^2(3)=2.00$, $\eta^2=.00$, $p=.572$), the highest completed level of education ($\chi^2(3)=3.08$, $\eta^2=.01$, $p=.380$), age ($\chi^2(3)=.25$, $\eta^2=.00$, $p=.969$), and English reading proficiency ($\chi^2(3)=.41$, $\eta^2=.00$, $p=.939$).

Table W2.4.2 reports operationalizations, literature sources, and scale reliability metrics for the content user perception experiment that we conducted using a survey.

Table W2.4.2: Operationalizations & Measures of Main Variables for Survey

Variable	Items	Source	Scale Reliability[1]
Readability	Bipolar 5-point scale with following items: “Please indicate whether you perceive the text above as … • poorly written – well written • poorly readable – well readable • not fitting together well – fitting together well • not understandable – understandable • not interesting – interesting”	Pitler and Nenkova 2008	.91
Understandability	Bipolar 5-point scale with following items: “Please indicate whether you perceive the text above as … • complicated – simple • unclear – clear • chaotic – orderly • illogically arranged – logically arranged • wordy – concise • difficult – easy“	Kamoen et al. 2013	.88
Credibility	Bipolar 5-point scale with following items: “Please indicate whether the text above is … • unbelievable – believable • inaccurate – accurate • not trustworthy – trustworthy • biased – not biased • incomplete – complete”	Roberts 2010, Flanigan and Metzger 2000	.87
Attitude toward the content	Bipolar 5-point scale with following items: “Please indicate whether you feel that the text above is … • distant – appealing • reluctant – inviting • boring – fascinating • impersonal – personal • monotonous – varied • interesting – uninteresting”	Kamoen et al. 2013	.89

[1]Cronbach’s Alpha with optimized number of items

In addition to the scales employed in W2.4.2, we measure content naturality using two items. On bipolar five-point scales, we ask respondents to indicate whether they believe that the content feels artificial vs. feels natural, and machine-made vs. human-made. We also ask two question to assess future intent. To gauge willingness to further inform, we use a slider from 0 to 100 and ask respondents to indicate how they agree with the statement: “I want to further inform myself about the company providing the service.” To measure willingness to buy, we use a slider from 0 to 100 and ask respondents to indicate how much they agree with the statement: “I am willing to buy the described service.”

2.4.3 Content Perception Inter-Correlations

Table W2.4.3 reports pairwise correlations between user perception variables using Kendall’s tau b, illustrating high correlations between these items.

Table W2.4.3: Consumer Content Perception: Dimensions’ Intercorrelations

	Kendall’s tau b (τ_b)						
Dimension	Readability	Understandability	Credibility	Attitude Toward the Content	Content Naturality	Willingness to Further Inform	Willingness to Buy
Readability	1.00**	.59**	.57**	.50**	.52**	.41**	.42**
Understandability		1.00**	.43**	.58**	.57**	.44**	.46**
Credibility			1.00**	.40**	.44**	.33**	.37**
Attitude Toward the Content				1.00**	.58**	.52**	.53**
Content Naturality					1.00**	.44**	.49**
Willingness to Further Inform						1.00**	.69**
Willingness to Buy							1.00**

[1]Statistical significance codes: *0.05 level, **0.01 level, one-tailed; n=551;

2.5 Consumer Content Perceptions: Computational Analyses

Table W2.5 illustrates computational analyses using LIWC (Pennebaker et al. 2015), the evaluative lexicon (Rocklage et al. 2018), and the text analyzer (Berger et al. 2020b) software packages that apply various lexica, analyses and scales to assess linguistic properties along psychological dimensions including concreteness, familiarity, and emotionality. The analysis reveals that differences between the semi-automated and human content are minor along most dimensions.

Table W2.5: Consumer Content Perception (Computational Analysis)

Dimension	Descriptives (Mean, SD)[1]				Kruskal Wallis[2]			
	Revised Machine	Real SEO Experts	Quasi Experts	Novices	χ^2	η^2	*df*	*p*
Concreteness	323.10 (7.45)	326.00 (5.37)	321.30 (7.48)	318.60 (4.28)	9.67	.15	3	.021*
Familiarity	574.14 (7.95)	578.14 (12.73)	579.22 (9.33)	581.47 (9.14)	7.14	.11	3	.067
Emotionality	3.28 (.66)	3.33 (.38)	3.47 (.55)	3.53 (.47)	3.07	.05	3	.380
Emotional Valence	6.15 (.89)	6.23 (.86)	6.45 (.77)	6.69 (.72)	3.70	.06	3	.296
Negations	.004 (.005)	.005 (.003)	.006 (.003)	.007 (.006)	3.28	.05	3	.351
Interrogatives	.011 (.006)	.009 (.004)	.013 (.006)	.013 (.008)	2.31	.04	3	.509
Causation	.028 (.009)	.030 (.013)	.032 (.015)	.026 (.009)	2.07	.03	3	.558
Certainty	.011 (.005)	.013 (.005)	.021 (.009)	.019 (.009)	16.24	.25	3	.001**
Tentativeness	.022 (.010)	.027 (.014)	.022 (.010)	.022 (.009)	1.44	.02	3	.697
Differentiation	.020 (.009)	.026 (.014)	.021 (.009)	.021 (.011)	1.25	.02	3	.740
Focus on future	.009 (.006)	.013 (.006)	.011 (.006)	.015 (.007)	8.54	.13	3	.036*

[1]Dimension scales: for concreteness, familiarity scale range: 100 (abstract, unfamiliar) to 700 (concrete, familiar), emotionality scale range: 0 (no emotion) to 9 (high emotion), emotional valence scale range: 0 (highly negative) to 9 (highly positive); other dimensions like negations, interrogatives, etc., represent percentages of total words in the text;
[2]Statistical significance codes: *0.05 level, **0.01 level; n=66;

2.6 Website Engagement

Having compared performance in terms of consumer perceptions and linguistic content, we next examine the impact of using semi-automated content on firm performance in terms of consumers' engagement with the website (e.g., Bronnenberg et al 2016, Jerath et al 2014, Edelman and Zhenyu 2016). We collect website traffic data for 412 days after the experimental content was posted. During this time, the content received 254 page views from 122 unique website visits arising from organic search results. Consistent with prior research, a series of χ^2 tests (e.g., Ghose et al. 2019, Azzopardi et al. 2018) reveal that semi-automated content performs better than human-generated content on the basis of the number of page views ($\chi^2(3)=257.31$, $p<.000$), page views from unique website visits ($\chi^2(3)=130.52$, $p<.000$), and the number of sessions started on the website through the SEO content (76, $\chi^2(3)=114.21$, $p<.000$). These results are consistent with the higher search engine rankings and the consumer search behavior that typically favors clicking on few, top ranked pages (Azzopardi et al. 2018). The semi-automated content also results in longer visits per visited page ($\chi^2(3)=167.15$, $p<.000$), suggesting better content performance (Danaher et al. 2006).

Based on a short survey for website visitors, we derive three proxies for expected performance: absolute buying affinity, relative buying affinity and expected sales. These metrics suggest that the semi-automated content offers superior performance to content crafted by SEO experts, beyond simply generating more page views due to its higher ranking, indicating a substantial positive financial impact on the company in the future. These results are presented in Table W2.6a.

Table W2.6a: User Behavior (Organic Search Source Only)

Dimension	Descriptives (∑)				One-Sample Chi-Squared[1]		
	Revised Machine	Real SEO Experts	Quasi Experts	Novices	χ^2	*df*	*p*
No. of Pages with Pageviews	16	3	5	10	11.88	3	.007**
No. of Pages with Pageviews in %	84.21	33.33	26.32	52.63	40.98	3	.000**
Pageviews	172	16	18	48	257.31	3	.000**
Unique Pageviews	84	6	9	23	130.52	3	.000**
Entrances	76	6	9	21	114.21	3	.000**
Exit Rate (means)	.41	.28	.32	.36	-	-	-
Bounce Rate	.00	.00	.00	.00	-	-	-
Avg. Usage Duration (Abs., sums)	3671	262	455	473	6639.40	3	.000**
Avg. Usage Duration (Rel.)[2]	229	87	91	47	167.15	3	.000**
Returning Visitors (Abs.)	88	10	9	25	127.09	3	.000**
Returning Visitors (Rel.)[2]	5.50	3.33	1.80	2.50	-	-	-
Buying Affinity (Abs.)[3]	4097	276	429	983	6670.60	3	.000**
Buying Affinity (Rel.)[2,4]	256	92	86	98	152.43	3	.000**
Exp. Sales (for U.P.*100)[5]	168	12	18	46	151.30	3	.000**

[1]Statistical significance codes: *0.05 level, **0.01 level;
[2](Rel.) = the absolute value (Abs.) divided by No_of_Pages_with_Pageviews
[3]Buying Affinity (Abs.) = Unique_Pageviews*Willingness_to_Buy (survey measured);
[4]Buying Affinity (Rel.) = Buying_Affnity (Abs.)/No._of_Pages_with_Pageviews;
[5]Exp. Sales (for U.P.*100) = (Unique_Pageviews/100*Expected_Sales_Rate)*100, where the expected sales rate is 2% (obtained from past company reports);

Table W2.6b reports statistics for the user behavior for visitors coming from direct links (e.g., links in emails, on other webpages, etc.) to the focal experimental pages on the website.

Table W2.6b: User Behavior (Direct Links Source Only)

Dimension	Descriptives (∑)				One-Sample Chi-Squared[1]		
	Revised Machine	Real SEO Experts	Quasi Experts	Novices	χ^2	*df*	*p*
No. of Pages with Pageviews	19	9	19	19	-	-	-
No. of Pages with Pageviews in %	100	100	100	100	-	-	-
Pageviews	545	126	257	515	342.65	3	.000**
Unique Pageviews	270	65	131	257	164.05	3	.000**
Entrances	226	47	95	222	166.57	3	.000**
Exit Rate (means)	.35	.35	.35	.35	-	-	-
Bounce Rate	.04	.07	.04	.04	-	-	-
Avg. Usage Duration (Abs., sums)	705	189	536	317	361.40	3	.000**
Avg. Usage Duration (Rel.)[2]	37	21	28	17	8.96	3	.029*
Returning Visitors (Abs.)	275	61	126	258	178.81	3	.000**
Returning Visitors (Rel.)[2]	14.47	6.77	6.63	13.57	-	-	-
Buying Affinity (Abs.)[3]	10021	3041	5846	12760	7066.10	3	.000**
Buying Affinity (Rel.)[2,4]	418	338	308	638	156.99	3	.000**
Exp. Sales (for U.P.*100)[5]	540	130	262	514	328.11	3	.000**

[1]Statistical significance codes: *0.05 level, **0.01 level;
[2](Rel.) = the absolute value (Abs.) divided by No_of_Pages_with_Pageviews
[3]Buying Affinity (Abs.) = Unique_Pageviews*Willingness_to_Buy (survey measured);
[4]Buying Affinity (Rel.) = Buying_Affnity (Abs.)/No._of_Pages_with_Pageviews;
[5]Exp. Sales (for U.P.*100) = (Unique_Pageviews/100*Expected_Sales_Rate)*100, where the expected sales rate is 2% (obtained from past company reports);

2.7 Content Production Costs Calculation Details

Table 4 in the main manuscript depicts the costs of content production and content machine induced savings. In this section, we elaborate on the calculations for Table 4. We took available working times and salary stats for the human reviser / SEO expert necessary (i.e., as stated in the table's footer: 39 hours available working time per week, 1,567 hours per year; 45,000 € of salary per year) and calculated the times, costs and possible output per year when using the manual way vs. the machine for text generation. To estimate the time spent per unit of content, information was provided by both the company and the experiment participants. Using this information, one can calculate theoretical outputs per year and labor costs per year. The calculation in Table 4 in the main manuscript are as follows:

- "Human labor time for content production" = empirically determined
- "Server cost per unit (€)" = empirically determined
- For human groups: "Produced content units" = "1,567 hours per year" / "Median (hours)"
- For machine to keep total costs at 45,000 € (i.e., the same as the human costs): "Produced content units" = 45,000 € / ("Labour cost per unit (€)" + "Server cost per unit (€)")
- Production level (%) = (("Produced content units Revised machine" / "Produced content units Company (real)")*100)-100
- "Labour cost per unit (€)" = 45,000 € / "Produced content units"
- "Cost for 164.95 units (€)" = ("Labour cost per unit (€)" + "Server cost per unit (€)")*164.95
- "Cost for 2,164.03 units (€)" = ("Labour cost per unit (€)" + "Server cost per unit (€)")*2,164.03
- "Produced content units (Possible real financial impact (2015 to 2019))" = empirically determined
- "Cost (€)" = 439*("Labour cost per unit (€)" + "Server cost per unit (€)")
- "Possible savings (€)" = "Cost (€)" of the Company (Real) – "Cost (€)" of specific comparison group

3 Education Sector Field Study

3.1 Keywords to Optimize for & Keyword Stats

Table W3.1 depicts the keywords (search queries) the experimental groups in the Education sector field study produced optimal content for, including keyword statistics (e.g., the average monthly search volume, i.e., how many people search on average per month for the keyword, the payed keyword competition, the keyword length), and descriptive statistics for the ranking performance of the revised machine content in the search engine. The keywords were selected such that they reflect target content and search queries in co-ordination with the educational institution running the experiments. The target keywords consist of short and long tail keywords, and mostly low search volume and competition keywords.

Table W3.1: Keywords for field experiments

Keyword	Avg. monthly search volume	Competition	Competition index	Keyword length	Mean revised machine ranking	Median revised machine ranking	SD revised machine ranking	IQR revised machine ranking	% of days revised machine was in ranking
	Descriptives								
quantitative marketing	10	low	-	2	6.30	5	4.80	1	100.00
quantitative marketing research	10	low	0	3	8.13	6.5	6.60	5.25	93.00
quantitative marketing program	-	-	-	3	2.03	2	0.16	0	88.40
marketing research seminar series	-	-	-	4	-	-	-	-	00.00
deep learning marketing	10	low	43	3	33.70	12	58.30	2	18.60
machine learning in marketing	10	low	16	4	84.10	98	39.10	64.5	100.00
machine learning in marketing education	-	-	-	5	1.00	1	0.00	0	34.90
digital marketing and machine learning	0	-	-	5	10.70	10	1.10	1	100.00
natural language processing in marketing	10	-	-	5	10.60	11	0.70	1	95.30
artificial intelligence in marketing	50	low	32	4	36.70	35	11.60	11.5	16.30
ai in marketing	10	low	30	3	29.00	29	0.00	0	1.00
ai in digital marketing	10	mid	36	4	77.10	78	5.70	8	67.40
ai in social media marketing	10	low	14	5	74.90	83	22.20	7.75	62.80
marketing with ai	10	-	-	3	32.50	27	22.90	14.5	46.50
marketing automation	320	mid	55	2	86.00	87	7.50	7.5	14.00
customer analytics	20	low	26	2	23.00	23	4.20	3	4.65
customer segmentation with machine learning	0	-	-	5	9.50	10	1.80	2	100.00
quantitative market research methods	10	-	-	4	8.70	10	4.20	8	86.00
business analytics education	10	-	-	3	1.00	1	0.00	0	1.00
career in marketing research	10	low	0	4	7.10	7	0.40	0	32.60
career opportunities in marketing	10	low	0	4	34.00	25	15.60	13.5	6.98
methods of marketing analytics	-	-	-	4	39.00	39	0.00	0	4.65
understanding digital marketing analytics	-	-	-	4	2.00	2	0.00	0	79.10
marketing phd career opportunities	-	-	-	4	2.00	2	0.60	0	58.10
quantitative marketing phd	10	-	-	3	4.60	1	9.00	1	93.00
doctorate PHD program in marketing	-	-	-	5	15.00	13	7.60	2.5	100.00
master program in marketing	10	mid	57	4	2.30	2	0.90	0	10.00
service marketing research	10	low	0	3	4.20	1	4.10	6	100.00
research in service marketing	-	-	-	4	4.70	5	2.30	3	100.00
academic research in service marketing	-	-	-	5	1.50	1	3.20	0	100.00
marketing institute college	-	-	-	3	10.60	3	20.30	2	100.00

Entries that display "-" mean that the search engine keyword tool did not provide specific information.

3.2 Supplemental Content Performance Tests

Table W3.2a reports group comparison tests for the search engine ranking performance of the experimental groups "revised machine" and "human" using Wilcoxon rank sum tests. Precisely, Table W3.2a shows that the revised machine outperforms the human content generating group in terms of the number of pages that got into the top10 search engine ranking, and in terms of mean ranking.

Table W3.2a: Search Engine Rankings Performance Comparison (Education Sector)

Dimension	Group	Descriptives					Wilcoxon rank sum [2]			
		n_p[1]	Median	(IQR)	Min	Max	W	z	*r*	*p*
Pages in ranking / day	**Revised Machine**	30	**18.00**	**(4.00)**	**12**	**22**	**908.00**	**-2.04**	**-.23**	**.041***
	Human		19.00	(2.00)	16	22				
Pages in top 10 / day	**Revised Machine**	30	**11.50**	**(3.00)**	**7.00**	**14.00**	**24.50**	**7.20**	**.82**	**.000****
	Human		5.00	(2.00)	3.00	9.00				
Mean rankings / day	**Revised Machine**	30	**17.57**	**(9.44)**	**5.60**	**30.12**	**1270.5**	**-5.79**	**-.66**	**.000****
	Human		26.22	(2.72)	19.44	30.30				

[1]n_p=number of pages per experimental group. n=77 (days); 2Two-tailed tests; statistical significance codes: *0.05 level, **0.01 level; Compared numbers are daily aggregate numbers.

Table W3.2b reports statistics for the experimental groups (e.g., Revised machine, Humans, etc.), the top 10 and the lowest ranked results on the 5 quality score components. The results are consistent with our findings from the IT service industry experiment.

Table W3.2b: Quality Score Components Group Comparisons to Top 10 Ranked Websites (Education Sector)

Quality Score Component	Group	Descriptives				Wilcoxon rank sum[1]			
		Median	(IQR)	Min	Max	W	z	*r*	*p*
Topic (s_a)	**Top 10**	**.44**	**(.20)**	**.15**	**.63**				
	Revised machine	.57	(.13)	.29	.76	205	3.72	.48	.000**
	Raw machine	.56	(.12)	.25	.76	223	3.42	.44	.000**
	Humans	.43	(.09)	.25	.64	476	-.37	-.05	.708
	Worst 10	.20	(.08)	.12	.38	803	-6.18	-.81	.000**
Keywords (s_k)	Top 10	**.47**	**(.23)**	**.09**	**.70**				
	Revised machine	.65	(.16)	.27	.85	194	3.90	.50	.000**
	Raw machine	.63	(.15)	.17	.84	190	3.97	.51	.000**
	Humans	.49	(.16)	.25	.73	393	.83	.11	.406
	Worst 10	.15	(.11)	.06	.42	805	-6.23	-.81	.000**
Uniqueness (s_d)	Top 10	**.95**	**(.06)**	**.78**	**.99**				
	Revised Machine	.94	(.08)	.73	1.00	527	-1.13	-.15	.258
	Raw machine	.90	(.09)	.72	.99	623	-2.57	-.33	.010*
	Humans	.99	(.05)	.07	1.00	234	3.19	.41	.001**
	Worst 10	.94	(.03)	.86	.98	459	.35	.05	.724
Readability similarity (s_r)	Top 10	**.57**	**(.07)**	**.46**	**1.00**				
	Revised Machine	.73	(.25)	.47	1.00	135.5	4.64	.60	.000**
	Raw machine	.78	(.23)	.45	1.00	85	5.39	.70	.000**
	Humans	.47	(.54)	.00	1.00	544.5	-1.39	-.18	.165
	Worst 10	.41	(.23)	.15	.70	726	-4.41	-.57	.000**
Naturality similarity (s_n)	Top 10	**.57**	**(.08)**	**.44**	**.75**				
	Revised Machine	.54	(.50)	.17	1.00	461	-0.15	-.02	.876
	Raw machine	.67	(.50)	.33	1.00	393	.84	.11	.403
	Humans	.17	(.25)	.00	1.00	780	-4.89	-.63	.000**
	Worst 10	.41	(.16)	.20	.88	725.5	-4.40	-.57	.000**

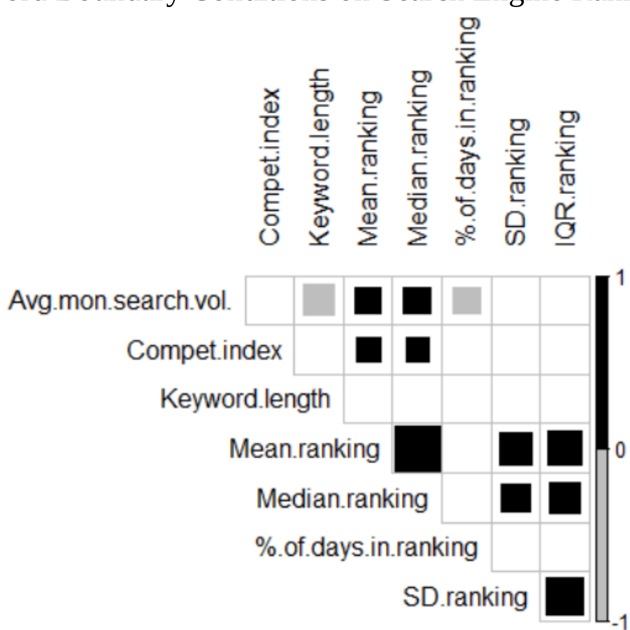
[1] Two-tailed tests; statistical significance codes: *0.05 level, **0.01 level;

4 Keyword Boundary Conditions

In this section, we investigate possible keyword boundary conditions (e.g., keyword length, competition levels, etc.) on the revised content machine search engine ranking performance using the data and keywords reported in Tables W2.1 (IT service sector) and W3.1 (education sector).

Figure W4 contains a series of Kendall's tau b correlations in which we try to ascertain if factors such as keyword length, search volume, or competition have an effect on the search engine ranking performance of the revised machine content. Figure W4 shows that the machine-generated content does not perform as well (i.e., higher values of mean and median search engine ranking positions) for more popular search terms (higher avg. monthly search volume and more competition). In addition, a higher avg. monthly search volume is also associated with a lower percentage of days for which machine-generated content is ranked in the search engine over our observation period. The keyword length does not play an important role.

Figure W4: Keyword Boundary Conditions on Search Engine Ranking Performance



Kendall's tau b correlations: ■ = positive correlation, ■ = negative correlation, □ = statistically non significant correlation (at a 0.05 alpha level), a bigger square represents a higher correlation coefficient;

5 Appendix References

Azzopardi L, Thomas P, Craswell N (2018) Measuring the utility of search engine result pages: an information foraging based measure. *SIGIR '18: The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval*. 605-614.

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, Lua JW, Kuha J, Lowe W, Müller C, Young L, Soroka S, Fellows I (2020) Quantitative analysis of textual data. Package 'quanteda'. Version 2.1.1. *CRAN*. Accessed May 20, 2019, https://cran.r-project.org/web/packages/quanteda/quanteda.pdf

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

Ghose A, Ipeirotis PG, Li B (2019) Modeling consumer footprints on search engines: An interplay with social media. *Management Science*. 65(3):1363-1385.

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.