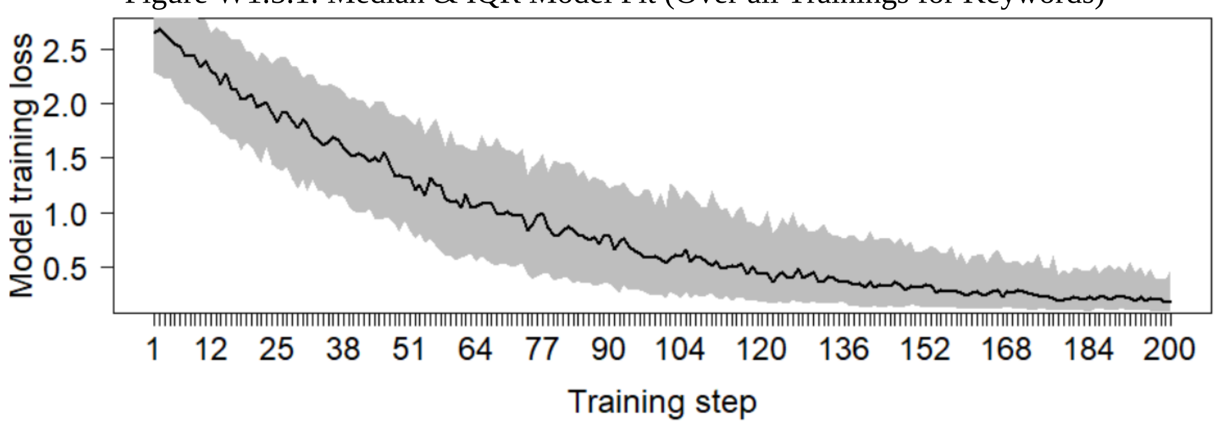
1.5 Validation of Method Fine-Tuning Process

For our experiments, we fine-tune our model for 200 training steps for each keyword, generating 100 pieces of content at each 20^{th} 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.

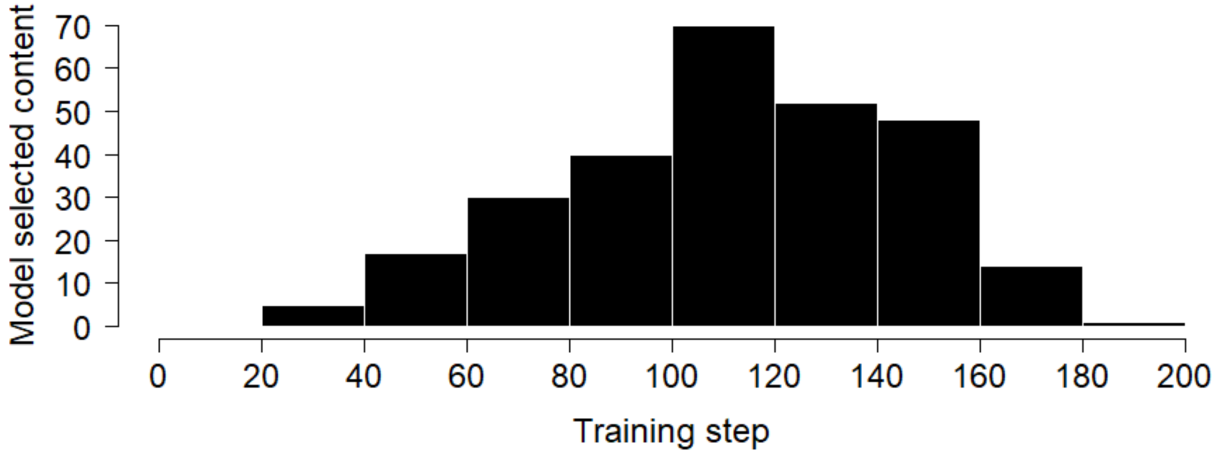
Using a robust regression (robust against violations of classic data assumptions of regression, see Maechler et al. 2020) for the training steps generating the "best" texts with highest overall quality scores on the quality score components, we observe in Table W1.5 that the content uniqueness among 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.

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.