2 IT Service Industry Field Study

This section contains details and accompanying materials for the empirical performance evaluation of our semi-automated content generation machine in the IT service industry application setting reported in the article.

2.1 Keywords Used & Keyword Statistics

Table W2.1 depicts the keywords (search queries) for which the experimental groups in the IT service sector field study produced content, including keyword statistics and descriptive statistics for the ranking performance of the revised machine content in the search engine. The basic keyword statistics reported in Table W2.1 include the average monthly search volume that serves as an indicator of how many users on average search for the keyword per month, the paid keyword competition to capture SEA competition that is provided by the search engine, and the keyword length to account for how many sub-words constitute the keyword. Our company partner selected the keywords used in the experiment based on its standard procedure for keyword selection (i.e., based on monthly search volume, competition, fit with the firm and keyword strategy), with a preference in favor of keywords with lower search volume in the long tail as part of their keyword strategy. The selected keywords are similar in terms of keyword statistics (competition, search volume, keyword length), and the company did not have any prior search engine ranking history for any of the keywords that they provided to us.

Table W2.1: Keywords for IT Service Field Experiment

Keyword	Descriptives								
	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
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

To generate website landing page content for the company we collaborated with, the participants of the three groups of human content writers described in the IT service industry application study in the article had access to all tools and the environment that the company uses

in its common SEO content production workflow. The instructions provided to the study participants are presented in Table W2.2.1. The participants were offered 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.

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. We controlled for writers' differences in terms of educational and other background covariates we considered relevant for the writing task. To this end we conducted a series of Kruskal Wallis group comparison tests. We find that the 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; $\chi^2(3)=5.89$, $\eta^2=.12$, $p=.053$); we also could not detect any significant differences in terms of time invested conducting research on the target keyword / topic ($\chi^2(3)=.28$, $\eta^2=.00$, $p=.868$) and content writing ($\chi^2(3)=3.76$, $\eta^2=.08$, $p=.153$).

Table W2.2.1: 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.

Descriptive statistics on the content length and changes are provided in Table W2.2.2. As these statistics show, the produced content is of about equal length across the experimental

groups and human revisers changed about 9% of the machine-made content in the revision process. We report evaluations of the performance of each of the experimental groups with regards to search engine rankings and the quality score in the next section of this document.

Table W2.2.2: 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).

2.3 Performance of the Experimental Groups

2.3.1 Quality Scores of Experimental Groups' Content

In Table W2.3.1, we compare the quality score components for the content generated by each of the experimental groups 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 terms of uniqueness. This finding demonstrates the content writing machine's capability to mimic patterns such as topics, keywords and readability levels contained in the top10 pages more thoroughly than human writers.

In comparing the experimental groups to the top 10 search engine rankings, bear in mind that the scores are comparative measures. For most of the components s_a, s_k, s_r and s_n, the top 10 ranked pages score lower than the raw and revised machine-generated content (and also some of the groups of human writers). This means that the machine-generated content is more similar along these dimensions to the set of the top ranked pages than the group of top 10 ranked pages are internally (i.e., comparing any given page from the top ranked pages to the set of top ranked pages). For example, consider the search query / keyword "IT service management". Content ranked on places 1-5 might focus about a relevant sub-topic (e.g., a telephone hotline for IT service management) of the keyword / search query, and content ranked on places 6-10 might focus on a different relevant sub-topic (e.g., a ticketing system for better customer support in IT service management) for the keyword / search query. The content writing machine is opportunistic in a sense, as it taps into features from all of the top 10 pages. That is, it picks, combines and rearranges features from all of the top 10 pages in a novel way. This "averaging" effect occurs during the fine-tuning process, which makes use of all of the top 10 pages. This "averaging" is taken into account in the quality score, in which all components compare a given content to the set of the top ranked content, which is the essence of SEO: trying to integrate all relevant features and aspects (that are present in the content of the top ranked pages) to fulfill the user's search query best (e.g., to provide the most rich information, formulated in the most appropriate way) (see section 1.5 in this document).

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

In Table 2 of the article we reported that the semi-automated content outperforms the remaining experimental groups in terms of (top 10) search engine rankings. To check robustness of this finding and to test whether it also generalizes for pairwise comparisons of experimental groups we conducted a series of Kruskal Nemenyi post hoc tests. Table W2.3.2 reports the results which show that the search engine performances of all experimental groups are statistically different at the 0.05 level.

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 SEO, content is usually optimized for a single main keyword (search query). However, in search engine advertising (SEA) ads and bids are often optimized for multiple keywords at once. Thus, a company could potentially benefit from optimizing SEO content for multiple keywords simultaneously. That is why we conducted an additional experiment for our IT service industry study to explore how well the experimental groups' SEO content ranks for related keywords. We identified the latter 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 based on 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 W2.3.3). 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.

Based on this auxiliary study, we conclude that the method seems to perform surprisingly well for related keywords as well. In contrast to relying on heuristics such as keyword density, the fine-tuning process of our semi-automated algorithm appears to not only capture the overall topic but also 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.

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

	Descriptives			
Group	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 top 10 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 W2.3.4 depicts the additional keywords for which content was generated. Consistent with our previously reported findings, Table W2.3.5 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 W2.3.4: 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 W2.3.5: 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 Providing Quality Score Feedback to Revise Content

To explore the incremental value of using NLG relative to SEO experts with access to the quality score further, we conducted the following additional study. Following the content that was initially produced by the real SEO experts we offered them the opportunity to improve their content by providing them with the quality scores of their initial content and the real search engine ranked top 10 content. The experimental setup mirrors A/B testing and is as follows: The tests were conducted using an online survey (including personal explanations of the task and a Q&A section). The survey contained the task description, the principal investigator's contact details, and an incentive of 40€ for the revision / feedback round per piece of content. The participants were introduced to the quality score and provided with an explanation of each quality score component and how to interpret it. They were provided with the quality score (both on each component and overall) for the content they produced initially. For comparison purposes, they were also provided with the top 10 ranked content for the specific keyword and their associated quality scores. Study participants entered their revised text 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 testing was conducted for 30 pieces of SEO expert produced content.

Table W2.4.1 shows that the SEO experts changed their original content by 10.24% (~77.50 words), ranging between 12 words changes and 176 word changes. Table W2.4.2 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 W2.4.1: Descriptives for Real SEO Experts Content Revision

Dimension	**Groups**	**Descriptives**			
		Median	(IQR)	Min	Max
Produced content length (in words)	Original SEO Expert Content	729.5	(47.25)	587	819
	Revised SEO Expert Content	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 SEO experts and revised SEO experts content such as added words, deleted words, and words with at least one changed letter (including changed letter capitalization).

Table W2.4.2: 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 SEO Experts	**.41**	**(.18)**	**.30**	**.72**				
	Revised SEO Experts	.44	(.18)	.30	.72	423	.39	.05	.697
Keywords (s_k)	Original SEO Experts	**.46**	**(.20)**	**.16**	**.79**				
	Revised SEO Experts	.46	(.15)	.16	.80	423	.39	.05	.697
Uniqueness (s_d)	Original SEO Experts	**.97**	**(.06)**	**.08**	**1.00**				
	Revised SEO Experts	.96	(.07)	.63	.99	560	-1.62	-.21	.105
Readability similarity (s_r)	Original SEO Experts	**.72**	**(.54)**	**.02**	**1.00**				
	Revised SEO Experts	.66	(.71)	.02	1.00	469	-.27	-.04	.784
Naturality similarity (s_n)	Original SEO Experts	**.67**	**(.39)**	**.08**	**1.00**				
	Revised 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;

2.5 Consumers' Content Perceptions

We provide details for the MTurk study presented in the article in which we examine differences in consumer perceptions between the semi-automated and human content. The instructions with which survey participants were presented are shown in Table W2.5.1.

Table W2.5.1: 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]

To assure data quality in our survey-based content 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' characteristics 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 W2.5.2 reports operationalizations, references, and scale reliability metrics for the content perception study we conducted.

Table W2.5.2: 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 scale items included in Table W2.5.2, we measure perceived 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 questions 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."

In Table W2.5.3, we report pairwise correlations between user perception variables using Kendall's tau b, illustrating high correlations between these items.

Table W2.5.3: Consumer Content Perception: Dimensions' Inter-Correlations

Dimension	**Kendall's tau b (τ_b)**						
	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;

We complement the MTurk Study with a computational linguistic analysis. 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, we assess the linguistic properties along psychological dimensions including concreteness, familiarity, and emotionality. The analysis presented in Table W2.5.4 reveals that differences between the semi-automated and human content are minor along most dimensions.

Table W2.5.4: 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. In addition to the findings discussed in the manuscript that indicate improved performance of semi-automated content relative to human-generated content on a number of dimensions (details are reported in Table W2.6.1), we also find that the semi-automated content results in longer visits per visited page ($\chi^2(3)=167.15$, $p<.000$), suggesting better content performance (Danaher et al. 2006).

Table W2.6.1: 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 W2.6.2 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 W2.6.2: 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 Cost Calculation Details

Table 4 in the article presents the cost of content production and savings induces by applying the content writing machine. In this section, we elaborate on the calculations for Table 4. We took available working times and salary statistics 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, cost 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. Based on this information, we can calculate expected outputs and labor cost per year. The calculation in Table 4 in the main manuscript are based on the following inputs, with the values from Table 4 appearing in quotations:

- "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 example, for the column "Company (real)" in Table 4: 1,567/9.5 ≈ 164.95
- 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 (€)"). For example, for the column "Revised Machine" in Table 4: 45,000/(15.79+5.00) ≈ 2,164.03
- "Production level (%)" = ("Produced content units" (e.g., of the revised machine) / "Produced content units" Company (real))*100-100. For example, for the column "Revised Machine" in Table 4: (2,164.03/164.95)*100-100 ≈ 1,211.95
- "Labor cost per unit (€)" = 45,000 € / "Produced content units". For example, for the column "Company (real)" in Table 4: 45,000/164.95 ≈ 272.81
- "Cost for 164.95 units (€)" = ("Labour cost per unit (€)" + "Server cost per unit (€)")*164.95. For example, for the column "Company (real)" in Table 4: 272.81+0*164.95 ≈ 45,000
- "Cost for 2,164.03 units (€)" = ("Labour cost per unit (€)" + "Server cost per unit (€)")*2,164.03. For example, for the column "Company (real)" in Table 4: 272.81+0*2,164.03 ≈ 590,369

- “Produced content units” (“Possible real financial impact (2015 to 2019)”) = empirically determined
- “Cost (€)” = 439*(“Labour cost per unit (€)” + “Server cost per unit (€)”). For example, for the column “Company (real)” in Table 4: 439*(272.81+0) ≈ 119,765
- “Possible savings (€)” = “Cost (€)” of the Company (Real) – “Cost (€)” of specific comparison group. For example, for the column “Revised Machine” in Table 4: 119,765-9,127 ≈ 110,638

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.