ChatGPT vs Human-authored Text Insights into Controllable Text Summarization and Sentence Style Transfer

Dongqi Pu Vera Demberg

Department of Computer Science
Department of Language Science and Technology
Saarland Informatics Campus, Saarland University, Germany

European Research Council
Established by the European Commission

Abstract

ChatGPT, a Generative Pre-trained Transformer model, has demonstrated remarkable capabilities in generating coherent answers given brief instruction prompts. The present study critically examines ChatGPT's performance in tasks involving controlled output variations for different audiences and writing styles. Our results highlight that human-authored texts exhibit more stylistic diversity than those produced by ChatGPT. Moreover, ChatGPT constantly incorporates inaccuracies while tailoring text to fit specific styles. These findings shed light on the **strengths** and **limitations** of ChatGPT.

Introduction

ChatGPT has shown promise in various natural language processing tasks. However, its efficacy in controllable text generation remains underexplored. In this study, we scrutinize ChatGPT's abilities in generating summaries for distinct audiences and altering sentence styles. We aim to answer: **How does the accuracy and style of ChatGPT-generated content differ from the human-produced text?**

Key contributions of this study include:

- Initial exploration of ChatGPT's effectiveness in controllable text generation.
- Highlighting performance disparities between ChatGPT and humans.
- Identifying and quantifying subtle errors in ChatGPT's text generation.

Experimental Setup

For all experiments, we utilized ChatGPT *gpt-3.5-turbo* with the following hyper-parameter setting: temperature = 0, top p = 1, frequency penalty = 0.2, and presence penalty = 0.2. Summary generation and sentence style transfer were constrained to a maximum of 512 and 32 generated tokens respectively. Default values were used for the remaining parameters.

Study on Controllable Text Summarization

Prompt version	FRE	CLI	DCR
layman	37.26†	14.82†	11.21†
simple	31.92†	15.70†	11.54†
simplified and understand.	35.48†	15.17†	11.21†
easy-to-comprehend	36.59†	14.93†	11.32†
straightforward	31.74†	15.58†	11.42†
general audience	35.86†	14.98†	10.96†
human answer (for layman)	53.06	12.36	8.90
expert	29.89†	15.91†	11.88†
technical	36.65†	13.76†	12.20†
comprehensive and detailed	31.62†	15.47†	11.15†
difficult-to-comprehend	28.95†	16.14†	11.71†
in-depth	34.37†	14.93†	10.82†
complicated	29.05†	15.76†	11.40†
human answer (for expert)	22.54	17.65	11.79

Table 1. Reading difficulty on different prompts, tested on a set of 500 randomly selected items. † indicates statistical significance ($p<0.05$) against corresponding human answers via paired t-test.

Candidate	FRE	CLI	DCR
Human Layman	52.42	12.46	8.93
Human Expert	23.20	17.62	11.78
ChatGPT Layman	37.38†‡	14.78†‡	11.17†‡
ChatGPT Expert	30.38†‡	15.82†‡	11.85†‡

Table 2. Reading difficulty scores by automatic metrics; † and ‡ indicate statistical significance ($p<0.05$) against same-style human answers, and opposite-style ChatGPT answers via paired t-test, respectively.

Table 1 suggests that different prompts for ChatGPT's layman summaries generally yield more readable than its expert summaries on the elife dataset. Further, as Table 2 demonstrates, ChatGPT effectively tailors summary complexity according to prompts across the entire dataset. However, the readability discrepancy is less stark than in human-created texts.

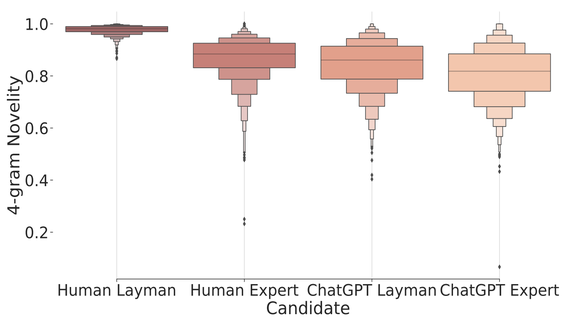

Figure 1. Comparison of abstractiveness between ChatGPT and human-generated summaries

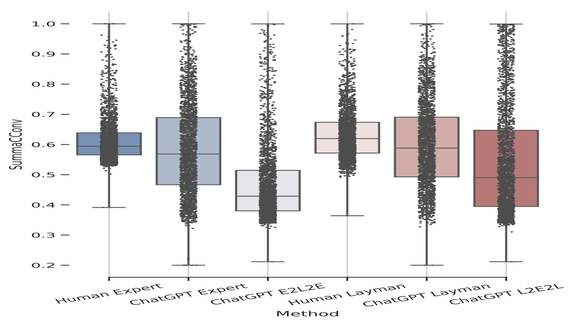

Figure 2. Summary consistency detection. L stands for layman, E for expert.

Candidate	Precision	Recall	F1
Human Layman	0.78	0.63	0.70
Human Expert	0.92	0.61	0.73
ChatGPT Layman	0.75‡	0.47†	0.58†
ChatGPT Expert	0.90‡	0.49†	0.63†
ChatGPT L2E2L	0.74‡	0.39†‡	0.51†‡
ChatGPT E2L2E	0.88‡	0.47†‡	0.62†‡

Table 3. Named entity hallucination on Elife dataset. † and ‡ indicate statistical significance ($p<0.05$) against same-style human answers, and opposite-style ChatGPT answers via paired t-test, respectively. L stands for layman, E for expert.

Our study reveals disparities between ChatGPT and human summarization in readability adaptability. ChatGPT's summaries lean more extractive (see Figure 1), showing a weaker correlation with human-authored summaries. Notably, misinformation (Figure 2) and hallucinations (Table 3) risks are higher in ChatGPT, with lower consistency scores and inaccurate handling of named entities.

Project Info

Paper

Study on Text Formality Transfer

Prompt version	Formality	MTLD
informal	51.09	13.22†
unprofessional	51.20	16.23†
spoken version	51.30†	14.47†
easygoing	51.43†	14.11†
casual	51.00	16.30†
laid-back	51.27	13.94†
human answer (for informal)	50.76	11.42
formal	52.22†	31.23†
professional	51.96†	31.98†
written	51.62†	29.69†
stately	51.30†	34.43†
grandiose	52.85†	30.71†
majestic	52.23†	33.49†
human answer (for formal)	53.92	14.99

Table 4. Text formality on different prompts, tested on a set of 500 randomly selected items. † indicates statistical significance ($p<0.05$) against corresponding human answers via paired t-test.

Dataset	Candidate	Formality	MTLD
GYAFC-FR	Human Informal	49.87	15.20
	Human Formal	53.57	18.70
	ChatGPT Informal	50.77†‡	14.60‡
	ChatGPT Formal	52.06†‡	31.68†‡
GYAFC-EM	Human Informal	50.11	12.11
	Human Formal	53.76	15.82
	ChatGPT Informal	51.02†‡	12.01‡
	ChatGPT Formal	51.98†‡	29.80†‡

Table 5. Text formality scores by automatic metrics; † and ‡ indicate statistical significance ($p<0.05$) against same-style human answers, and opposite-style ChatGPT answers via paired t-test, respectively.

ChatGPT can effectively modify sentence formality (Table 4 and 5), leaning towards higher formality levels and showing greater lexical diversity in formal text, likely due to bias in its training data.

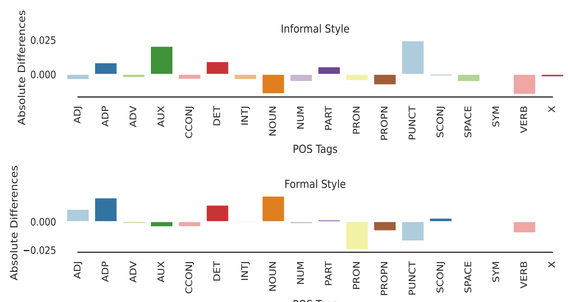

Figure 3. Absolute differences in POS tags distribution of ChatGPT and human-generated sentences: GYAFC - EM

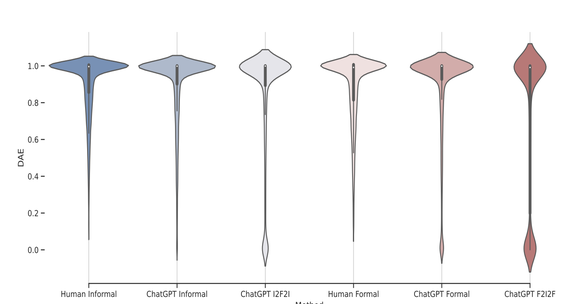

Figure 4. Dependency arc entailment: GYAFC - EM. Data points>0.95≈Accurate. To clarify discrepancies, cutoff point=0.95.

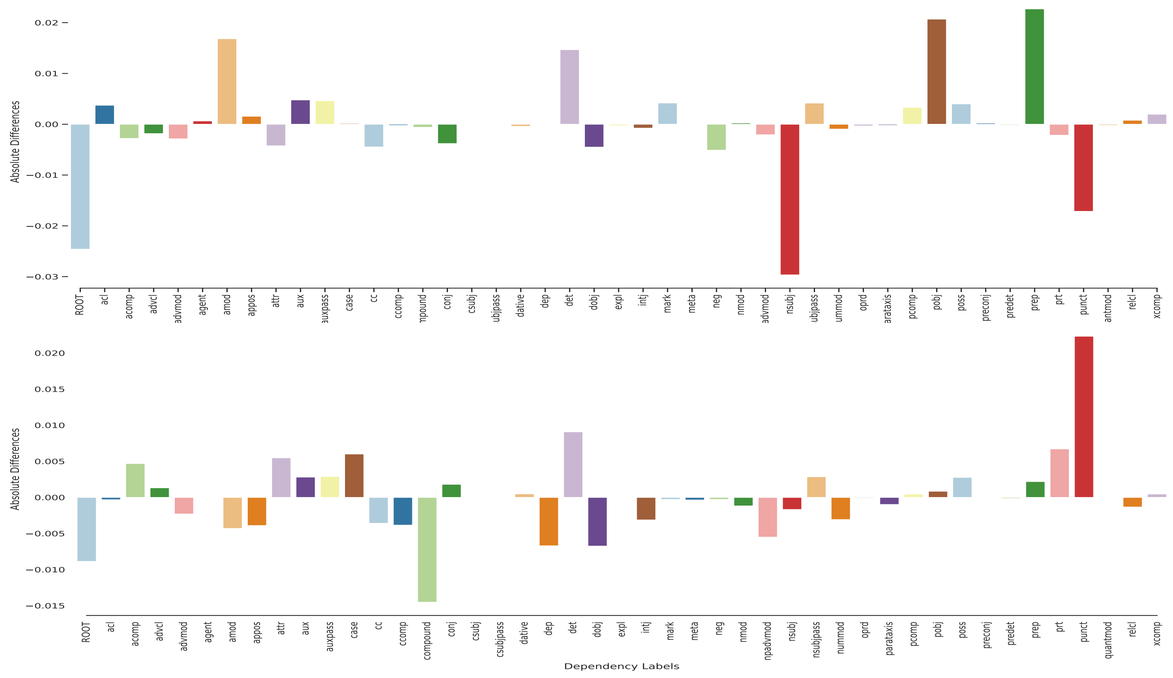

Figure 5. Absolute differences in dependency labels distributions of ChatGPT and human-generated sentences: Upper-Formal, Lower-Informal

ChatGPT's formal and informal style control differs significantly from human patterns, with variations in POS tags usage and dependency labels. Moreover, when ChatGPT undergoes multiple text transformations, it exhibits an increased risk of factual inconsistencies and hallucinations, further deviating from human performance.

Conclusion

In this study, we provide an extensive evaluation of ChatGPT's text-generation ability. We find that there are notable differences from human-authored texts. Our research also reaffirms concerns about hallucinations and inaccuracies within ChatGPT's outputs.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878).

My concurrent work at ACL 2023

Incorporating Distributions of Discourse Structure for Long Document Abstractive Summarization (ACL 2023, Main Conference)

https://dongqi.me Saarland University, 66123, Germany dongqi.me@gmail.com