Evaluating Text Summarization Generated by Popular AI Tools

Zhuoran Lin, Sifan Chen, Ning Wang and Hongjun Li
China Agricultural University, Beijing, China

Keywords: Evaluation, Text Summarization, Large Language Models (LLMs), Jensen-Shannon Divergence.

Abstract: Automatic summarization is a crucial component of Natural Language Processing and has long been a prominent area of research. This study focuses on evaluating the performance of several well-known AI tools, namely ChatGPT, Claude, Bart, Pegasus, and T5-Base_GNAD, in the field of text summarization. To conduct the evaluation, we assembled a corpus comprising fifty abstracts from various subject fields. The Jensen-Shannon divergence (DJS) metric was employed to assess the accuracy of these models. The findings indicate the following: a) Bart outperforms other AI models in the task of summarization, with ChatGPT3.5 and Pegasus following closely behind, b) ChatGPT3.5 demonstrates proficiency in Agricultural Science. Bart's summarization capabilities are more evenly distributed. Notably, in the domain of physics, all AI tools yield relatively higher DJS scores, while performing well in Arts & Humanities and Interdisciplinary subjects. c) Statistical significance tests conducted between the models reveal substantial differences, and both ChatGPT3.5 and Bart exhibit significant performance variations across subject fields.

1 INTRODUCTION

Since OpenAI, a San Francisco based AI research lab, released their product -- ChatGPT3.5, it has become immensely popular. Within days of launching, ChatGPT3.5 attracted hundreds of thousands of users who were fascinated by its ability to conduct natural conversations.

ChatGPT3.5 demonstrates how advanced AI has become at Natural Language Processing (NLP). It can understand complex sentences, keep context in mind, and respond appropriately by generating coherent and fluent responses. This ability to have engaging back-and-forth conversations has captured public interest in ChatGPT. Many people find chatting with ChatGPT to be an amusing or intriguing experience, even though it's "just" an AI system. The enthusiasm for ChatGPT demonstrates why continuing progress in natural language AI is so important and eagerly anticipated.

As far as we know, a massive amount of researchers have conducted extensive researches on AI tool, especially in ChatGPT from Dec. 2022 to Apri. 2023. One of the most solid and thorough overall capability assessment of ChatGPT is conducted by OpneAI (2023) itself, which shows ChatGPT4.0 has overall capability in text understanding, generation. Some scholars used different kinds of standard examinations to assess the competence of ChatGPT (Sarah W. Li, Fares Antaki). Also there are researches focus on the application of ChatGPT in different industries and make an assessment for it (Sarah W. Li, Brent J. Sinclair, Enkelejda Kasneci). Several studies focus on specific ability. Sun et al. (Sun Hao, 2023) conducted a safety assessment of Chinese Large Language Models (LLM) comparing several AI tool (e.g. ChatGPT, BELLE BLLOM).

This paper specifically concentrates on the capability of text summarization of ChatGPT3.5 and comparing it with other popular AI tools. There are another four AI tools (Claude from Slack, Bart from Meta, Pegasus from Google, T5).

1.1 Models Introduction

(1) Claude is an artificial intelligence chatbot developed by Slack using machine learning algorithms trained on massive datasets.
(2) The Bart model (Lewis Mike, 2019) was initially pre-trained on the English language and subsequently fine-tuned using CNN Daily Mail data. It demonstrates remarkable efficacy when fine-tuned for various text generation tasks, such as summarization and translation. Additionally, it exhibits strong performance in comprehension tasks, including text classification and question answering.
(3) Pegasus employs a pretraining task intentionally

Lin, Z., Chen, S., Wang, N. and Li, H.
Evaluating Text Summarization Generated by Popular AI Tools.
DOI: 10.5220/0012283800003807

In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 350-354
ISBN: 978-989-758-677-4

designed to resemble summarization. In this task, crucial sentences are deliberately removed or masked from an input document, and the model generates these sentences as a single output sequence alongside the remaining sentences. This approach is akin to producing an extractive summary (Zhang Jingqing, 2020). Furthermore, Pegasus has demonstrated state-of-the-art performance in summarization across all 12 downstream tasks, as assessed by metrics like ROUGE and human evaluations.

(4) T5-Base_GNAD is a fine-tuned variant that has attained the subsequent results on the evaluation set: Loss (2.1025), Rouge-1(27.5357), Rouge-2 (8.5623), Rouge-L (19.1508), Rougelsum (23.9029), and Generation Length (52.7253).

Automatic summarization stands as a pivotal challenge within the field of NLP, presenting numerous complexities encompassing language comprehension (such as discerning the vital content components) and content generation (including the aggregation and rephrasing of identified content to produce a summary) (Sreyan Ghosh, 2022).

When categorizing types of summaries, we encompass two dimensions: extractive summarization and abstractive summarization. Extractive summarization entails the creation of a summary that is a subset of the original text, as it contains all the words present in the original text, while abstractive summarization potentially contains new phrases and sentences that may not appear in the source text.

To our best knowledge, the vast majority of researchers used text database (like Wikipedia, CNN/DM, TAC dataset and so on) to evaluate capability of information extraction of LLM (Li Liuqing, Cabrera-Diego), but rarely researchers perform a evaluation of text summarization of ChatGPT and other AI tools (e.g. Claude) or language models by using different subject materials (e.g. Agricultural Science, Physic, Chemical, Computer Science). We have selected ten subject fields based on Web of Science (WOS) Categories and then collected five highly cited theses in each field, chosen by peer researchers for their high-quality abstracts and analytical content.

2 METHODOLOGY

This paper aims to conduct research utilizing English abstracts from various fields, ranging from Agricultural Science to Philosophy & Religion. Our metric of choice is the Jensen-Shannon divergence (D_{JS}), which has exhibited strong correlation with manual evaluation methods such as Pyramid, Coverage, and Responsiveness, in predicting system rankings (Louis Annie, Saggion H.).

Before delving into the details of the Jensen-Shannon divergence (JS divergence), it is important to introduce the concept of Kullback-Leibler divergence (KL divergence) (Kullback S. 1951). KL divergence is an information-theoretic measure that quantifies the dissimilarity between two probability distributions over the same event space. Within information theory, KL divergence can be interpreted as a measure of information loss when multiple messages are encoded using a second distribution. In the context of summary evaluation, this translates to encoding a source document using an Automatic Text Summary (ATS) system. Consider two probability distributions, P and Q, where P represents the distribution of words in the source document and Q represents the distribution in the candidate summary. The Kullback-Leibler (KL) divergence is defined as follows:

$$D_{KL}(P \| Q) = \frac{1}{2} \sum_W P_W \log_2 \frac{Pw}{Qw} \quad (1)$$

While the resulting values of KL divergence are always non-negative, it lacks the symmetric property ($D_{KL}(P||Q) \neq D_{KL}(Q||P)$), fails to satisfy the triangular inequality, and tends to yield divergent values (Thomas M. Cover, 2012). To address these limitations, Lin et al. (Lin C.Y. 2006) proposed the use of the Jensen-Shannon divergence (D_{JS}) to measure information loss between two documents. The DJS is formally defined by Equation (2):

$$D_{JS}(P \| Q) = \frac{1}{2}\left(\sum_W P_w \log_2 \frac{2Pw}{Pw+Qw} + Q_w \log_2 \frac{2Qw}{Pw+Qw}\right) \quad (2)$$

In Equation (2), P_w represents the probability distribution of term win the source document, while Q_w represents the probability distribution of term w in the candidate summary. The probability distribution of each term w is computed using Equation (3):

$$\mathcal{P}(\omega) = \frac{C + \delta}{N + \delta * B} \quad (3)$$

Here *C* is the count of word w and *N* is the number of tokens. Specifically, we set $\delta = 1*10^{-10}$ and B = |V|, where V stands for the number of all different terms obtained from source document and candidate summary.

Based on this evaluation method, we applied our corpora to the AI models, allowing them to summarize the abstract texts. Specifically, we instructed the models to generate a summary containing approximately n words (n = the number of words multiplied by 30%) to prevent excessive rephrasing. We collected all the generated summaries and utilized Python to calculate the D_{JS} scores.

Since certain language models (LLMs) lacked specific websites with chat-box interfaces, we

downloaded some models from the hugging face repository, which hosts a vast collection of LLMs. To illustrate our data processing methodology, we provide two examples. Example 1 demonstrates one of the samples used in our testing. Each time we input a prompt, such as "Summarize the following paragraph using n words," the chat-bot generates an answer similar to the response in the Bot1 box. We repeated this process, collecting the generated summaries and incorporating them into our dataset. Example 2 showcases a Python code snippet used to input our corpora and employ the LLMs downloaded from hugging face to generate summaries.

Example1

User: summarize the following paragraph within n words:

The demand for food is expected to significantly increase with continued population growth over the next 50 years, ...（227 words）

Bot1: Population growth drives increased food demand. Mulching and nitrogen fertilizers impact soil environment and crop yield for food security....(227*30% words)

Example2

```
from transformers import pipeline
import pandas as pd
import csv
summarizer = pipeline("summarization", model="facebook/bart-large-cnn")
df = pd.read_excel('/Users/*****/Desktop/****.xlsx')
csvf = open('bart.csv','a+',newline='')
writer =csv.writer(csvf,delimiter=',')
for ARTICLE in df('abstraction'):
n = len(ARTICLE.split(‘ ’))
text_list = ()
  text = summarizer(ARTICLE, max_length=n*0.3, min_length=n-15, do_sample=False)(0)('summary_text')
text_list.append(text)
  writer.writerow(text_list)
csvf.close()
```

3 RESULTS

3.1 Comparison of Models

Bart demonstrates the highest summarization capability among the five models, achieving a score of 0.315 (Table 1). This score significantly surpasses those of the other four models. T5-Base_GNAD and Claude obtain scores of 0.383 and 0.393, respectively, while ChatGPT3.5 and Pegasus achieve scores of 0.347 and 0.357. The differences between the former two models and the latter two models are statistically significant. Consequently, the summarization abilities of these five models can be categorized into three levels. Bart exhibits the highest proficiency, followed by ChatGPT3.5 and Pegasus in the intermediate range, while T5-Base_GNAD and Claude demonstrate relatively weaker performance in the task of summarization.

3.2 Comparison of Subjects

Based on the comprehensive analysis presented in Table 1, it is evident that the performance of the five models varies across different subject fields. In Physics, all models exhibit relatively higher D_{JS} scores, 0.385 in average, indicating their poor proficiency in summarizing abstracts from this field. However, their performance excels in the Arts & Humanities, Interdisciplinary field, with average score of 0.338.

When focusing on specific subject fields, it is worth noting that ChatGPT3.5 and Claude struggle in text summarization tasks related to Physics, both surpassing a D_{JS} score of 0.4. In contrast, Bart performs exceptionally well in Arts & Humanities, Interdisciplinary subjects, achieving a score of 0.295. However, Claude consistently lags behind other models, demonstrating lower performance across various subject fields.

In the field of Agricultural Science, ChatGPT3.5 obtains the lowest score of 0.310, while both Claude and T5-Base_GNAD rank last, scoring 0.397. In Biology & Biochemistry and Chemistry, Bart emerges as the best performer, while T5-Base_GNAD performs the poorest. Similarly, in Clinical Medicine and Computer Science, Bart remains the top-performing AI tool, while Claude struggles, scoring above 0.4.

The deviation is huge in Psychiatry/Psychology, with Bart scoring the lowest (0.281) and Claude obtaining the highest score (0.404). Mathematics proves to be a challenging subject for all five models, except Bart, as their scores range from 0.360 to 0.398. In Philosophy & Religion, Bart excels with a score of 0.280, while T5-Base_GNAD lags behind with a score of 0.404.

Across subject fields, significant differences in performance are observed between ChatGPT3.5 and Bart models. Notably, the performance of ChatGPT3.5 in Physics is significantly higher compared to Agricultural Science and Philosophy & Religion. Similarly, Bart demonstrates higher significance levels in Chemistry and Physics when

compared to Philosophy & Religion and Psychiatry/Psychology. Others show no significantly difference.

Table 1. D_{JS} score of various AI models in 10 different subjects.

Models	D_{JS} score										
	All subjects	Agricultural Sciences	Arts & Humanities, Interdisciplinary	Biology & Biochemistry	Chemistry	Clinical Medicine	Computer Science	Physics	Psychiatry/Psychology	Mathematics	Philosophy & Religion
ChatGPT3.5	0.347 B	0.310 a	0.342 ab	0.341 ab	0.349 ab	0.334 ab	0.353 ab	0.406 b	0.349 ab	0.360 ab	0.327 a
Claude	0.393 C	0.397 a	0.378 a	0.372 a	0.364 a	0.408 a	0.403 a	0.406 a	0.404 a	0.398 a	0.395 a
Bart	0.315 A	0.326 ab	0.295 ab	0.313 ab	0.345 b	0.328 ab	0.315 ab	0.340 b	0.281 a	0.327 ab	0.280 a
Pegasus	0.357 B	0.357 a	0.344 a	0.366 a	0.360 a	0.361 a	0.367 a	0.385 a	0.320 a	0.373 a	0.338 a
T5-Base_GNAD	0.383 C	0.397 a	0.332 a	0.404 a	0.392 a	0.395 a	0.374 a	0.390 a	0.364 a	0.377 a	0.404 a
Average	0.359	0.357	0.338	0.359	0.362	0.365	0.362	0.385	0.344	0.367	0.349

Note: The capital letter in column "All subjects" represents the difference significant among models, while the lowercase letter in same line shows the difference significant of each model's performance in different fields.

4 CONCLUSION AND FUTURE WORKS

This paper presents a comprehensive evaluation of Large Language Models (LLMs) in the context of text summarization. Our study employed a diverse corpus comprising abstracts from ten different subject fields. The results indicate that the Bart Model emerges as the most effective tool for text summarization tasks, achieving a score of 0.315, followed closely by ChatGPT3.5 with a score of 0.347, Pegasus with a score of 0.357, T5-Base_GNAD with a score of 0.383, and Claude with a score of 0.393.

In terms of subject-specific performance, ChatGPT3.5 demonstrates notable proficiency in Agricultural Science, but its performance in Physics is notably weak. Bart exhibits a well-balanced performance across subject fields, particularly excelling in Philosophy & Religion. Pegasus performs well in Psychiatry/Psychology but shows limitations in Physics, scoring 0.320 and 0.385, respectively. T5-Base_GNAD performs well in Arts & Humanities and Interdisciplinary subjects, but struggles in the fields of Biology & Biochemistry and Philosophy & Religion, both scoring over 0.4. Claude gets relatively well in Chemistry, Biology & Biochemistry, Arts & Humanities, Interdisciplinary subjects, while performing less effectively in other subject fields with D_{JS} scores around 0.4.

It is worth noting that AI models obtained relatively higher D_{JS} scores in the field of physics but excelled in Arts & Humanities and Interdisciplinary subjects. Significance testing revealed high levels of statistical significance among the five models, with specific significant differences observed between ChatGPT3.5 and Bart in certain subject fields. The order of statistical significance among the five models, from highest to lowest, is the group (Claude, T5), the group (ChatGPT3.5, Pegasus), and Bart.

To ensure the robustness of our findings, it is imperative to expand the scope of our research in future works. One crucial aspect that requires attention is the enlargement of our dataset to encompass a broader range of subject fields. By including a more diverse array of disciplines, such as Economics, Sociology, Political Science, and Engineering, we can obtain a more comprehensive understanding of the performance of Large Language Models (LLMs) in text summarization tasks across various domains.

In addition to expanding the dataset, we recognize the importance of employing more sophisticated evaluation metrics to enhance the accuracy and depth of our analysis. While the current evaluation primarily focuses on the D_{JS} scores, future work will incorporate additional metrics to evaluate the quality of the generated summaries. Coherence, which measures the logical flow and organization of the summary, and cohesion, which assesses the connectivity and smooth transition between ideas, will provide valuable insights into the structural integrity of the summaries.

Furthermore, evaluating the grammatical range and accuracy of the summarization output will allow us to gauge the linguistic proficiency of the LLMs. This aspect is crucial in ensuring that the generated summaries not only capture the essence of the source documents but also adhere to the grammatical rules and conventions of the target language. By incorporating these metrics, we can obtain a more nuanced understanding of the strengths and weaknesses of LLMs in the context of text summarization.

To achieve these objectives, future research will involve collaboration with domain experts and linguists to develop a comprehensive evaluation framework that encompasses a broader range of metrics. Moreover, the inclusion of human evaluators and professional linguists to assess the quality of the summaries will provide valuable insights and serve as a reliable reference for comparison. By expanding our dataset and employing more sophisticated evaluation metrics, we can enhance the accuracy, reliability, and validity of our research findings, ultimately contributing to advancements in the field of text summarization and the effective utilization of LLMs.

REFERENCES

Sarah W. Li, Matthew W. Kemp, Susan J.S. Logan, et al. ChatGPT Outscored Human Candidates in a Virtual Objective Structured Clinical Examination (OSCE) in Obstetrics and Gynecology(J). 2023. American Journal of Obstetrics and Gynecology. https://doi.org/10.1016/j.ajog.2023.04.020

Fares Antaki, Samir Touma, Daniel Milad, et al. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings (J). Ophthalmology Science. 2023, 3(4): 100324. https://doi.org/10.1016/j.xops.2023.100324

Brent J. Sinclair. Letting ChatGPT do your science is fraudulent (and a bad idea), but AI-generated text can enhance inclusiveness in publishing (J). Current Research in Insect Science. 2023, 3: 10057. https://doi.org/10.1016/j.cris.2023.100057

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, et al. ChatGPT for good? On opportunities and challenges of large language models for education (J). Learning and Individual Differences. 2023, 103: 102274. https://doi.org/10.1016/j.lindif.2023.102274

Sun Hao, Zhang Zhenxin, Deng Jiawen, et al. Safety Assessment of Chinese Large Language Models (Z). arXiv. 2023. https://arxiv.org/abs/2304.10436

Lewis Mike, Liu Yinhan, Goyal Naman, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (Z). arXiv. 2019. https://doi.org/10.48550/arXiv.1910.13461

Zhang Jingqing, Zhao Yao, Saleh Mohammad, et al. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv(Z). 2020. https://doi.org/10.48550/arXiv.1912.08777

Sreyan Ghosh. Abstractive Summarization with Hugging Face Transformers (EB/OL). Keras, 2022. https://keras.io/examples/nlp/t5_hf_summarization/

Li Liuqing, Jack Geissinger, William A. Ingram, et al. Teaching Natural Language Processing through Big Data Text Summarization with Problem-Based Learning(J). Data and Information Management. 2020, 4(1):18–43. https://doi.org/10.2478/dim-2020-0003

Cabrera-Diego, Luis Adrián, and Juan-Manuel Torres-Moreno. SummTriver: A New Trivergent Model to Evaluate Summaries Automatically without Human References (J). Data & Knowledge Engineering. 2018, 113:184–97. https://doi:.org/10.1016/j.datak.2017.09.001.

Louis Annie, Ani Nenkova. Automatically Evaluating Content Selection in Summarization without Human Models(C). In Conference on Empirical Methods in Natural Language Processing, 2009, 306–314. https://repository.upenn.edu/entities/publication/45db321d-f0bb-448d-9bb5-618c872e14c6

Saggion H., Torres-Moreno J. M., da Cunha I., SanJuan E., et al. Multilingual summarization evaluation without human models (C). In Coling 2010: Posters. 2010, 1059-1067.

Kullback S., Leibler R.A. On information and sufficiency (J). Annals of Mathematical Statistics. 1951, 22 (1): 79-86.

Thomas M. Cover, Joy A. Thomas. Elements of information theory (2nd ed.)(C). New York: John Wiley & Sons, Inc, 2012.

Lin C.Y., Cao G., Gao J., & Nie J. An information-theoretic approach to automatic evaluation of summaries(C). In Proceedings of the main conference on human language technology conference of the north American chapter of the association of computational linguistics. 2006, 463–470. https://aclanthology.org/N06-1059