# Prompt scoring system for dialogue summarization using GPT-3

George P. Prodan and Elena Pelican

*Abstract*—**Recent results in language processing show that language models are capable of performing several natural language tasks without the need of supervised learning. A challenging task for pre-trained language models is dialogue summarization. One way of generating summaries is engineering prompt templates for few-shot training. However, a static approach of creating prompts leads to unreliable outcomes between different classes of dialogues. Focusing on the dialogues structure properties, we propose a scoring system to improve the few-shot training performances by building tuned prompts composed by the highest scored dialogue samples. Our evaluation based on ROUGE scores and human evaluation shows that there is improvement for the experiments where we made use of the score system. The positive results are validated by all the three large-scale datasets we used in testing. All experiments were performed within the framework of the GPT-3 API.**

*Index Terms*—**natural language processing, pre-trained language models, abstractive summarization, prompt tuning.**

## I. Introduction

LANGUAGE models have evolved significantly in recent years. Whereas task specific models show that they can attain very good results only in one direction [1] [2], language models prove that they can handle a variety of NLP tasks without supervised learning. Colossal models such as GPT-3 are already used by thousands of developers. Few-shot learning without weights updating [3] is one reason for that, as it makes possible a fast development of applications in several directions (classification, semantic search, content generation, summarization and so forth). GPT series are based on the Transformer architecture [4], which relies on self-attention mechanisms [5] [6].

Dialogue summarization is a challenging problem that can be tackled using state-of-the-art NLP technologies. In this study we analyze the possibility to use such models for abstractive summarization of conversations. The aim is to perform generic and informative single-document summarization [7]. This can be done using GPT-3 in a few-shot setting, meaning that one constructs a prompt consisting of one or more summarized dialogues along with the input dialogue, which is given to the model for completion. The fact that this can be done only by using a few training samples, without changing any of the model weights, has several advantages over the classic fine-tuning: (i) there is no need in creating another model, so we save memory space and time and (ii) even if there is a lack of corpora for fine-tuning purposes, we can still obtain reliable results based only on prompt tuning.

Online communication platforms and mobile chat applications can implement functionalities based on dialogue summarization. Multiple notifications can be replaced by a few summaries providing a better user experience. There is a massive activity on chat applications nowadays, hence such a feature would match the needs of the users. However, one should take into account that processing an enormous volume of messages using models like GPT-3 will demand large computational costs. This leads to the problem of cost minimization. Another objective of this work is to study the behaviour of summarization performance when we are using less computational resources.

### A. Related work

Improving the few-shot training performances is a problem of interest as there can be instabilities in the GPT-3 performances due to the way prompt is chosen [8]. A solution based on contextual calibration was proposed for tasks such as text classification, fact retrieval or information extraction [9]. Dialogue summarization prompt-tuning techniques based on negation understanding and name substitution for dialogues with multiple participants are investigated in [10].

There have been investigated fine-tuning approaches which also prove that prompt-based tuning increases the language models performances [11] [12]. It is known that GPT series of models are multitask learners [13], but different tasks demand different prompt tuning approaches. Our focus is to find the best way of tuning the prompts for dialogue summarization. The research in dialogue summarization has started to gain popularity in the last two years [14]. There is an impressive increase of data-sets for the summarization of the chat conversations [15]–[18] and several models have been developed [19]–[28].

The main known problem of the dialogue summarization is when the summary provides wrong references [14] [29], distorting the information from the original dialogue, phenomenon also known as *hallucinations* [30], [31].

### B. Solution details

We aim to improve the quality of the summaries generated in a few-shot training regime by choosing the best picks for the training samples. For that, we establish a simple but efficient scoring system which takes into account the dialogue content, size and number of active participants in the conversation (Section 3) in order to find feature similiarities between

G. P. Prodan is with University of Padova, Italy and Ovidius University of Constanta, Romania.

E. Pelican is with Ovidius University of Constanta, Romania.

Manuscript received september, 2021; revised april, 2022.

dialogues. Two similar conversations will have a higher score than two different ones.

Firstly, we calibrate the scoring system configuration in order to achieve the best performances (Section 4). We separate each component and prove its relevance to the system. The scoring system is initialised with a weight distribution based on the initial experiments. The final distribution is established after several variations.

Secondly, we vary the model temperature. We noticed an increase of the performance around 25% for low temperatures. We continued the experiments at a low temperature of 0.25. More precise determinations can be done for each parameter we consider in our evaluations. However, at the moment there is no proof to state that the scoring system configuration we use will behave at its best for any other data-set. Further work will be needed to analyze that.

### C. Evaluations

The performance evaluation is based on the ROUGE scores [32]. The established score system is evaluated with respect to the results obtained with random generated prompts (Section 5). We observe an improvement of the ROUGE scores for the summaries generated with the score system (SS) proposed by us. Apart from evaluating the score system, we also tested the summarization behaviour for different GPT engines. We noticed that SS improves the summarization performances for each engine. Lastly, we performed human evaluation and compared the results with those obtained with ROUGE-1 and ROUGE-L F1 metrics.

## II. Experimental setup

We used the GPT-3 API[1] provided by OpenAI [3] to access the GPT-3 engines (text-curie-001, curie, curie-instruct-beta, ada, babbage).

### A. Datasets

The data-sets on which we rely in experiments are briefly described. They are large datasets that could be used in the training of the dialogue summarization:

1) **SAMSum Corpus** - SCd, *a Human-annotated Dialogue Dataset for Abstractive Summarization* [15] is a dataset published by Samsung researchers in 2019. It contains 16369 messenger conversations created by linguists. It is currently the largest provider of conversations similar to those discussed on online chats.
2) **DialogSum** - DSd, *a Real-Life Scenario Dialogue Summarization Dataset* [16] is similar to SAMSum Corpus, but spoken conversations are summarized instead. There are 13460 summarized dialogues labeled manually.
3) **MediaSum** - MSd, *a large-scale media interview dataset* [18] including a multitude of interview transcripts and their abstractive summaries. We use a part of this dataset (12460 transcripts that have up to 16 utterances).

### B. Implementation details

A preliminary processing takes place for the dialogues considered in the prompt shots selection - the selection pool (SP). We construct the term frequency matrix (TF) and compute the inverse document frequency coefficients (IDF). Also, we retrieve the token count and the number of persons participating in the conversation for each dialogue in SP.

In our experiments SP consists in 12k - 14k dialogues depending on the dataset used. The tests are taken one by one using input dialogues from the testing part of each dataset. The data reduction of the SP is performed only once before testing. For each SP dialogue we compute and save the features mentioned earlier (TF, IDF, token count, attendance and ID).

The score system is retrieving information about the SP dialogues only by searching through the SP features. The input dialogue which has to be summarized is analysed and the score system is assigning a score to it by comparing its features to each entry in the selection pool. Then, we use the IDs of the highest k scored dialogues in order to retrieve the content and the summary. Finally, we construct the prompt by following the order imposed by the scores and generate the summary using the GPT-3 engine. The prompt is engineered in such a way the summary to correspond with the prompt completion. For testing, we repeat this procedure for 50, 100 or 200 runs of the model in the same configuration.

The prompt includes the following components: an instruction hint, the selected samples and the delimiters. We provide an example for a two-shot prompt in Appendix A.

## III. Methods for scoring

Three aspects are considered in order to have the best picks of dialogues for the few-shot training of the model: the content, size and attendance. Each of them is characterised by a corresponding score and weight. When evaluating the score system performance, we are looking to find optimal configurations that lead high ROUGE scores [32] of the generated summaries.

### A. Content

The content evaluation is based on the TF-IDF approach [33] [34]. We rely on this method as it can be easily implemented on different SPs. Different variants of this approach can be further developed depending on the context and SP behaviour. Using the BERT tokenizer from transformers library [35], we determine the tokens distribution for each dialogue and computed the TF-IDF weights,

$$w_{td} = \frac{f_{td}}{f_{md}} \log \frac{N}{n_t}$$

where $f_{td}$ is the token $t$ frequency in dialogue $d$, $f_{md}$ is the maximum frequency of a token in the dialogue $d$, $N$ is the total number of dialogues and $n_t$ is the number of dialogues in which we can find token $t$.

The content score, $x_c$, is given by the cosine similarity between an input dialogue and a dialogue from SP.

[1] https://beta.openai.com/

### B. Size

We aim to find dialogues having a similar length to that of the input dialogue. Let $\mathcal{S} = A \times d_0$ be the set of dialogue pairs from which we search the training shots. For each selected pair $(d, d_0)$ we compute a length similarity coefficient $x_c$. This coefficient is given by a function $f$ which should obey the following conditions:

- $f : \mathcal{S} \to [0, 1]$
- $|d| = |d_0| \Leftrightarrow f(d, d_0) = 1, \forall d \in A$
- $\forall d_r, d_l$ such that $|d_r| - |d_0| = |d_0| - |d_l| > 0 \Rightarrow f(d_l, d_0) \geq f(d_r, d_0)$

The last condition increases the chances to choose the shorter conversation if there are multiple ones with similar content. There is no reason to increase the number of tokens in the query which is handled to the model if the content score is the same. From the distribution of the token count (Figure 1), one can observe that the short conversations are much more and the probability to find content similarities are high. Thus, in most of the cases we can perform the search only through a pool of shorter conversations.

The maximum score is achieved when we pick a conversation of exactly the same token count. If such a sample is not found, then we try to find another one with a small error. Such as the case of a standard deviation, there can be defined an interval within the samples to be scored high. Indeed, we can choose a standard distribution for scoring, meaning that we got a curve similar to a symmetrical Gaussian,

$$x_s = f(d, d_0) = const \times e^{-\frac{|d|-|d_0|}{\sigma}} \tag{1}$$

but which does not help if we are looking to minimize the costs. Shorter conversations help in creating a low token count prompt. Therefore, we need an asymmetrical curve in order to boost the probability of picking short dialogues. In this work we test the performance of the asymmetrical double sigmoid (*ads*) function in two configurations: narrow and broad. Different *ads* trends can be obtained by modifying the free parameters $c_1$-$c_6$,

$$\begin{aligned} f(d, d_0) =& c_1 + \frac{c_2}{1 + \exp\left(-\frac{|d|-|d_0|-c_3+c_4}{c_5}\right)} \\ & \times \left[1 - \frac{1}{1 + \exp\left(-\frac{|d|-|d_0|-c_3-c_4}{c_6}\right)}\right] \end{aligned} \tag{2}$$

The large number of free parameters allowed us to test different shapes and find a good set-up for our experiments. Average token count (ATC) of the sampled dialogues is the quantity we use to monitor computational costs, namely

$$ATC = \frac{1}{k}\sum_{i=1}^{k} |d_i|$$

where $|d_i|$ is the token count of a dialogue sampled from SP and $k$ the number of shots used in prompt tuning.

The function which computes the size coefficient has a significant contribution in cost minimisation. We test three functions: the normal distribution with $\sigma = 20$ tokens, meaning a FWHM of 47.2 tokens (gauss), an asymmetrical double sigmoid having a FWHM of 61.8 tokens (ads narrow), and another asymmetrical double sigmoid having a larger FWHM of 85.6 tokens (ads broad). We find that a broader *ads* function helps for cost minimisation. Also, the performance is not affected significantly. There are very small differences between the ROUGE-1 scores:

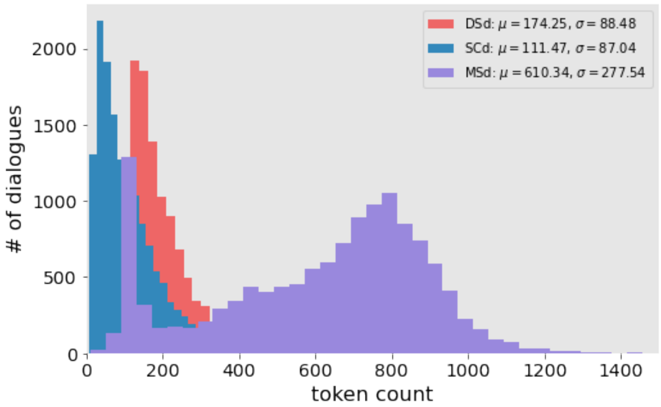


Fig. 1. Datasets statistics on token counts

- *random picking*: 228.04 average token count, 40.9 ROUGE-1 F1 average score
- *normal distribution* ($\sigma = 20$ tokens): 229.63 average token count, 41.76 ROUGE-1 F1 average score
- *ads narrow* (FWHM 61.8 tokens): 220.76 average token count, 41.84 ROUGE-1 F1 average score
- *ads broad* (FWHM 85.6 tokens): 215.96 average token count, 41.9 ROUGE-1 F1 average score

Obviously, there are many other options for the scoring function. If we are looking only to minimize the costs, then we can choose a function for which $f(d, d_0) = 0$ for any $d$ and $d_0$ such that $|d| > |d_0|$. Moreover, in order to increase the scoring for shorter conversations, we can give up on the condition that the score is maximum when the dialogues have equal token counts. In these cases, obtaining the best performance is not the top priority and the trade-off between costs and performance is evident.

We study the behaviour of two functions: the piecewise function,

$$f(d, d_0) = \begin{cases} 1 & \text{if } |d| < |d_0| \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and the following function, that can improve the chance in picking similar size conversations depending on the value of the free parameter $a$,

$$f(d, d_0) = \begin{cases} 1 + a(|d| - |d_0|), \text{if } |d| < |d_0| \\ 0, \text{otherwise} \end{cases} \tag{4}$$

We run 100 two-shots SCd tests for these functions using SS on 5:3:2 configuration, at the temperature of 0.25:

- *Piecewise function*: 119.88 average token count, 39.41 ROUGE-1 F1 average score
- *Equation* 4 ($a = 0.005$): 202.67 average token count, 41.63 ROUGE-1 F1 average score

TABLE I
EXPERIMENTAL RESULTS - WEIGHTS CONFIGURATION 200 TESTS

| $w_c$ | $w_s$ | $w_a$ | SCd<br>$\mu = 111.47$<br>$\sigma = 87.04$ | | | DSd<br>$\mu = 174.25$<br>$\sigma = 88.48$ | | | MSd<br>$\mu = 610.34$<br>$\sigma = 277.54$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | R-1 | R-2 | R-L | R-1 | R-2 | R-L | R-1 | R-2 | R-L |
| 0.1 | 0.5 | 0.4 | 39.88 | 15.54 | 31.27 | 36.7 | 13.55 | 29.2 | **39.04** | **19.18** | **29.81** |
| 0.1 | 0.6 | 0.3 | 39.68 | 15.73 | 31.37 | 37.05 | 13.38 | 29.6 | 38.13 | 18.43 | 28.87 |
| 0.1 | 0.3 | 0.6 | 39.71 | 15.45 | 31.07 | 36.68 | 13.21 | 29.25 | 38.25 | 18.79 | 29.23 |
| 0.5 | 0.3 | 0.2 | 39.83 | 15.2 | 31.51 | **37.51** | **13.8** | **29.7** | 37.89 | 18.43 | 28.41 |
| 0.3 | 0.4 | 0.3 | **40.38** | **16.07** | **31.83** | 37.26 | 13.62 | 29.36 | 37.51 | 17.93 | 28.46 |
| 0.6 | 0.3 | 0.1 | 39.8 | 15.07 | 31.15 | 37.22 | 13.61 | 29.67 | 38.14 | 18.34 | 28.54 |
| N/A<br>random picks | | | 39.69 | 15.03 | 31.36 | 34.88 | 11.35 | 27.73 | 32.81 | 13.42 | 23.98 |

### C. Attendance

The number of participants in a dialogue is also a key factor. As also suggested in [36], better results are obtained when the summarization system takes into account the number of persons involved in the discussion. The conversations between two people are usually easier to summarize than those when there are more participants involved. We investigate two scoring methods. In the first method, we group the conversations of the SP in two classes: 2 participants class and 3+ participants class. The attendance score is simply

$$x_a = \begin{cases} 1 & \text{if the dialogues are in the same class} \\ 0 & \text{otherwise} \end{cases}$$

The other method is based on the fact that the score should be maximum when the two dialogues which are compared have exactly the same number of participants. The score is decreasing based on the relative difference in the number of the participants, $|\Delta a|/a_0$. The attendance score can be calculated by using an exponential,

$$x_a = e^{-|\Delta a|/a_0}$$

In this way we get a consistent score in the interval [0, 1] and we penalise an arbitrary difference in attendance less if the reference dialogue has a larger number of participants.

We run 100 tests (2 shots, 0 temperature) for each method:

- *Method 1 (piecewise)*: 37.99 ROUGE-1 F1 average score
- *Method 2 (exp)*: 38.21 ROUGE-1 F1 average score

As method 2 leads to better performances, we use this in further experiments.

### D. Weights

The final score is given by

$$s = \langle w, x \rangle = w_c x_c + w_s x_s + w_a x_a$$

In creating the tuned prompt, we place the scored dialogues in an ascending order. The last shot should be the one scored highest as it has more impact on the completion returned by the model.

Several weights are tested whereas keeping the same GPT-3 configuration. By GPT-3 configuration, we refer only to the engine, temperature and number of shots. We use mainly the most recent version of curie engine, text-curie-001. The temperature is a measure of how random the tokens are chosen in the generated summary. A low temperature implies picking the most probable tokens predicted by the model. Thus, a higher temperature produces larger fluctuations in the results we obtain, and implicitly in the ROUGE scores. When we experiment on different weights we use a 0 temperature configuration to avoid fluctuations.

Preliminary experiments are done for 50 tests of GPT-3 text-curie-001 at a temperature of 0 to spot the best set of weights. We cannot decide on a single configuration of the weights as the tests are not numerous enough to arrive to a statistically significant conclusion. The experimental results are shown in Appendix B.

Further experiments are done on a smaller set of weights, by running 200 tests on each configuration. The results are presented in Table III-B. In the last row we show the scores obtained by using random prompts (no feature scoring is used). For all three datasets it is observed an increase in ROUGE scores when we use sample scoring. The effect of SS is clearly seen for the datasets containing longer dialogues (DSd and MSd). The average token counts of the datasets, $\mu$, along with their standard deviations, $\sigma$, are provided in the same Table III-B. For MSd, where there is a larger $\sigma$, we can also notice larger fluctuations in ROUGE scores and a tendency of obtaining better results when $w_s$ (size scoring weight) is larger. For the first two datasets, scoring based on content similarity ($w_c$) contributes more in choosing the best prompt.

## IV. Results and discussions

### A. Temperature

We vary the temperature using three SS configurations: SCd with $w_c = 0.5$, $w_s = 0.3$ and $w_a = 0.2$ (SCd-5:3:2), SCd with $w_c = 0.3$, $w_s = 0.4$ and $w_a = 0.3$ (SCd-3:4:3) and DSd $w_c = 0.5$, $w_s = 0.3$ and $w_a = 0.2$ (DSd-5:3:2). We run text-curie-001 on 100 two-shot prompts. In Figure 2 can be observed how we obtain better ROUGE scores for lower temperatures. Therefore, we conduct the experiments at a low temperatures setting (0.00 - 0.25).

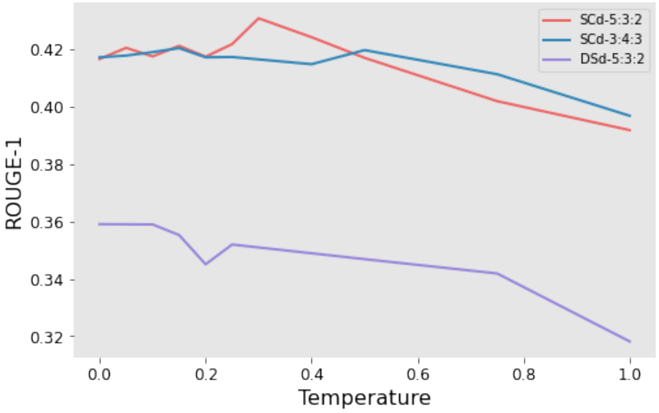


Fig. 2. Performance dependence on the model temperature. A number of 100 tests were performed for three configurations.

### B. Performance evaluation of the score system

We experimented further by running at different numbers of prompt tuning shots. As expected, the performance increases almost everytime when more tuning samples are provided to the prompt. Figure 3 shows the ROUGE-1 score as a function of the shots number. We repeated the experiment for each dataset.

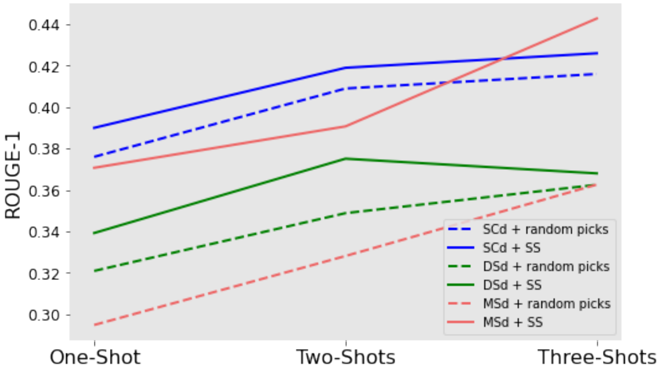


Fig. 3. Score system performance evaluations with respect to random generated prompts baseline. The weights configurations are those highlighted in Table I.

### C. Comparison to fine-tuning

We fine-tune a GPT-3 curie model for 4 epochs to compare its performance to the prompt-tuning method we propose. Different SPs between 50 and 2000 SCd samples are used for training the fine-tuned model. The results illustrated in Figure 4 show that fine-tuning clearly outperforms our prompt-tuning with 0.07 ROUGE-1 score units on average. Also, fine-tuning can demand more computational resources. For 4 epochs, training requires spending between 45k and 1600k tokens depending on the number of samples in SP (the training set), whereas, in this experiment, only prompt-tuning is spending on average 221 tokens per query. Thus, around 400 prompt-tuning based queries can run using the same amount of tokens elapsed in training the model on 100 samples. However, this can be a good investment if the fine-tuned model is used for a long-term period in a stable environment (meaning that the SP samples are still efficient as training examples after a certain amount of time). Using prompt-tuning or fine-tuning is a decision that should be analyzed depending on the targeted performance, costs and the expected amount of queries that the model will process.

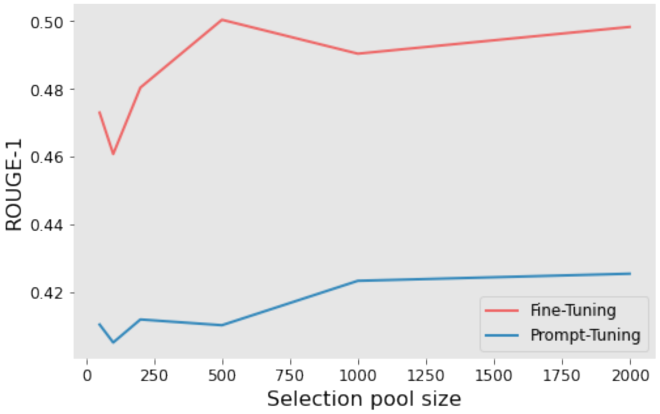


Fig. 4. Comparison between Prompt-Tuning and Fine-Tuning performances with respect to the selection pool size.

### D. Performance on smaller models

Using a smaller model can be useful when one needs to consume less resources. We evaluate the performances for ada, babbage, curie and curie-instruct-beta engines of GPT-3 with and without using SS. As expected, the results (Figure 5) show that overall performance decreases significantly when using a smaller model. However, using the score system is still an efficient way of improving the prompts.

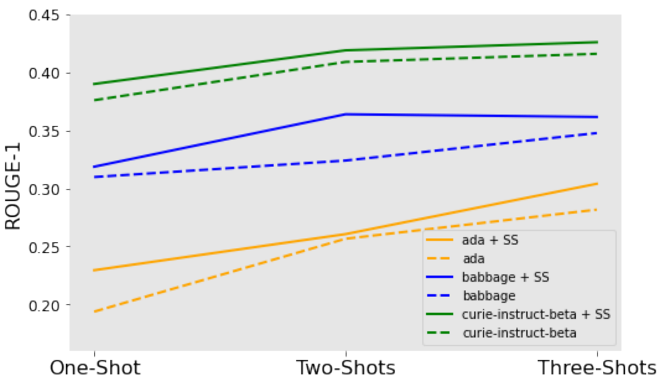


Fig. 5. Performance evaluation by ROUGE-1 scores for different GPT-3 engines. *Score system (SS): SCd-5:3:2 ads broad.*

### *E. Selection pool variations*

The SP for each dataset remains unchanged during the experiments. However, to investigate if the score system results depend on the SP we perform a series of separate experiments including a significant amount of foreign data.

We mix data from SCd and DSd to create a heterogeneous SP. The new SP consists of 54% samples from SCd (14732), 46% from DSd (12460). We run the model again for 100 tests with curie-instruct-beta engine at temperature of 0.25 and 5:3:2 ads-broad score system. Also, we run the model only on tests from DSd. The results are shown in Table II. As expected, in the case of one-shot training, the performance is slightly better for a larger SP. However, the SCd only two-shots training scores higher than for the other SPs. Also, there is a large difference in the score provided by SS between DSd experiments and the others. This difference is actually a measure of the discrepancies between SCd and DSd.

Software applications can be developed using the score system we propose. To do so, one must consider a way of defining the SP. Different categories of users may use different SP. Depending on the users behaviour, SP will also have to change in time. Consequently, a solution based on dynamic SP can be implemented. Users feedback also can be used in upgrading the SP in two ways: (i) by providing better summaries and (ii) by setting up several preferences. Methods for gathering data to improve or form new SPs could be a continuation of this work.

### *F. Human evaluation*

For human evaluation (HE) we consider four criteria following [37]: Coherence - it is related to the overall quality of the sentences and how well is the summary structured, Consistency - it measures the factual information transfer, Fluency - it scores the quality of individual sentences (i.e. grammar, formatting and so forth) and Relevance - it shows how important is the selected information, an excess of information is penalised. We ask the annotators to rate different summaries for 100 dialogues on a scale from 1 to 5. The evaluation is blind and it includes the reference summary from SAMSum dataset and the summaries generated by GPT-3's curie-instruct-beta (two shots and 5:3:2 low temperature configuration) with and without applying the score system. The results are presented in Table IV. The average scores of 3 x 100 dialogue summaries are provided together with the number of dialogues rated between unit intervals. A significant increase due to the use of the scoring system can be observed. For Relevance we can spot the largest difference in the average score. However, an improvement is visible for each criterion. The number of failures (i.e. poor summaries rated under 2.5) decreases with 11% of the total amount of dialogues after applying the scoring system.

We calculate the Pearson coefficient to evaluate the correlation between human judgment and ROUGE scores (Table V). We see that the ROUGE scores are not reliable in this case. There are many examples, as the one provided in Table III, when very good summaries are rated with low ROUGE scores.

## V. Conclusion

In this study we investigated possible prompt-based improvements of language models to perform abstractive summarization. We show that choosing the right dialogues can increase the quality of the summaries and reduce the number of failures by 11 %. We test the scoring system for several GPT-3 engines and we obtain better results for each engine when applying the scoring system. Also, the scoring system we proposed for selecting the best picks to create prompts also can control the computational costs by using different size similarity functions. We proved that content similarities between dialogues are also valuable for tuning the prompt, and as a result computational resources can be saved.

We evaluate the scoring system using ROUGE metrics and, also, by conducting human evaluation. It seems that in our experiment small variations in the average ROUGE score correspond to larger discrepancies in the scores given by the annotators. However, both evaluations show that applying the score system increases the quality of the summaries.

By studying the selection pool behaviour, we identified another research direction that can represent a future work. The way one gathers the dialogue samples in the selection pool can lead to different results. It may be necessary to engineer dynamic data-sets of dialogue samples depending on the end users behaviour and individual feedback. Thus, investigating prompt tuning methods that do not require fine-tuning is an advantage as we do not need to use space for distinct modules of a fine-tuned version of the model for each user.

## Acknowledgment

The authors would like to thank OpenAI for allowing us to access GPT-3 API for academic purposes. We gratefully thank the anonymous referees for their insightful comments and recommendations. We also thank the annotators for their contribution in evaluating the summaries.

## References

[1] J. Chai and A. Li, "Deep learning in natural language processing: A state-of-the-art survey," in *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2019, pp. 1–6.

TABLE II
THE ROUGE SCORES, AVERAGE SCORES PROVIDED BY OUR SCORE SYSTEM AND AVERAGE TOKEN COUNT (ATC) FOR DIFFERENT SELECTION POOLS AND NUMBER OF SHOTS USED IN THE TRAINING.

| SP | Training | R-1 | R-L | $s = \langle w, x \rangle$ | ATC |
|---|---|---|---|---|---|
| SCd | One-Shot | 39.0 | 31.5 | 0.645 | 109.40 |
| SCd + DSd | One-Shot | 40.8 | 32.3 | 0.6529 | 110.55 |
| DSd | One-Shot | 40.4 | 32.1 | 0.4635 | 118.59 |
| SCd | Two-Shot | 41.9 | 33.4 | 0.6255 | 215.96 |
| SCd + DSd | Two-Shot | 41.5 | 33.1 | 0.6335 | 216.34 |
| DSd | Two-Shot | 39.1 | 31.6 | 0.4377 | 242.62 |

TABLE III
EXAMPLE OF SUMMARY SCORED LOW BY ROUGE METRICS, BUT EVALUATED AS VERY GOOD BY HUMAN JUDGMENT

| **Dialogue** | Will: hey babe, what do you want for dinner tonight?<br>Emma: gah, don't even worry about it tonight<br>Will: what do you mean? everything ok?<br>Emma: not really, but it's ok, don't worry about cooking though, I'm not hungry<br>Will: Well what time will you be home?<br>Emma: soon, hopefully<br>Will: you sure? Maybe you want me to pick you up?<br>Emma: no no it's alright. I'll be home soon, i'll tell you when I get home.<br>Will: Alright, love you.<br>Emma: love you too. | | | |
|---|---|---|---|---|
| **Source** | **Summary** | **R1** | **RL** | **HE** |
| GPT-3 2-shot + SS | Emma is not feeling well and doesn't want to cook.<br>Will is trying to make sure she's ok. | 25.80 | 19.35 | **4.25** |
| SAMSum | Emma will be home soon and she will let Will know. | 100 | 100 | 3 |
| GPT-3 2-shot | Emma is not hungry and is not sure when she will be home.<br>Will is worried about her. | **48.27** | **34.48** | **4.25** |

TABLE IV
HUMAN EVALUATION RESULTS

| **Statistics** | SAMSum | **GPT-3 + SS** | GPT-3 random |
|---|---|---|---|
| Coherence | 4.23 | 3.87 | 3.58 |
| Consistency | 4.11 | 3.53 | 3.22 |
| Fluency | 4.16 | 3.86 | 3.61 |
| Relevance | 3.97 | 3.29 | 2.80 |
| Average Score | 4.12 | 3.64 | 3.30 |
| Between 4-5 | 68 | 47 | 37 |
| Between 3-4 | 26 | 32 | 34 |
| Between 2-3 | 15 | 25 | 32 |
| Between 1-2 | 4 | 10 | 19 |
| Failures (1 - 2.5) | 9 | 22 | 33 |

TABLE V
PEARSON CORRELATION COEFFICIENTS BETWEEN HUMAN EVALUATION RESULTS AND ROUGE METRICS FOR TWO SET-UPS: GPT-3 WITH SCORING SYSTEM AND WITHOUT IT (RANDOM PROMPTS)

| **Set-up** | **Criterion** | **R1** | **RL** |
|---|---|---|---|
| **GPT-3 + SS** | Coherence | 0.3324 | 0.3122 |
| | Consistency | 0.2860 | 0.2719 |
| | Fluency | 0.2328 | 0.2454 |
| | Relevance | 0.3665 | 0.3284 |
| | **Average** | **0.3478** | **0.3296** |
| **GPT-3 random** | Coherence | 0.1423 | 0.2052 |
| | Consistency | 0.0884 | 0.1415 |
| | Fluency | 0.1853 | 0.2332 |
| | Relevance | 0.1693 | 0.2499 |
| | **Average** | **0.1683** | **0.2391** |

[2] P. Kłosowski, "Deep learning for natural language processing and language modelling," in *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2018, pp. 223–228.

[3] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[4] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.

[5] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[6] Q. Guo *et al.*, "Low-rank and locality constrained self-attention for sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, p. 2213–2222, Dec. 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2944078

[7] "Abstractive summarization: An overview of the state of the art," *Expert Systems with Applications*, vol. 121, pp. 49–65, 2019.

[8] Y. Lu *et al.*, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity," *arXiv e-prints*, p. arXiv: 2104.08786, 2021.

[9] T. Z. Zhao *et al.*, "Calibrate Before Use: Improving Few-Shot Performance of Language Models," *arXiv e-prints*, p. arXiv:2102.09690, Feb. 2021.

[10] M. Khalifa, M. Ballesteros, and K. McKeown, "A bag of tricks for dialogue summarization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8014–8022. [Online]. Available: https://aclanthology.org/2021.emnlp-main.631

[11] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *ACL/IJCNLP*, 2021.

[12] S. Hu *et al.*, "Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification," *arXiv e-prints*, p. arXiv:2108.02035, Aug. 2021.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[14] X. Feng, X. Feng, and B. Qin, "A survey on dialogue summarization: Recent advances and new frontiers," *arXiv e-prints*, p. arXiv:1711.01731, 2021.

[15] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "Samsum corpus: A human-annotated dialogue dataset for abstractive summarization," *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 2019. [Online]. Available: http://dx.doi.org/10.18653/v1/D19-5409

[16] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, "Dialogsum: A real-life scenario dialogue summarization dataset," in *FINDINGS*, 2021.

[17] L. Mehnaz *et al.*, "GupShup: An Annotated Corpus for Abstractive Summarization of Open-Domain Code-Switched Conversations," *arXiv e-prints*, p. arXiv:2104.08578, Apr. 2021.

[18] C. Zhu, Y. Liu, J. Mei, and M. Zeng, "Mediasum: A large-scale media interview dataset for dialogue summarization," *arXiv preprint arXiv:2103.06410*, 2021.

[19] L. Zhao, W. Xu, and J. Guo, "Improving abstractive dialogue summarization with graph structures and topic words," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 437–449. [Online]. Available: https://aclanthology.org/2020.coling-main.39

[20] X. Feng *et al.*, "Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks," in *Chinese Computational Linguistics*. Cham: Springer International Publishing, 2021, pp. 127–142.

[21] ——, "Language model as an annotator: Exploring dialogpt for dialogue summarization," *ACL 2021*, May 2021.

[22] W. Chien-Sheng *et al.*, "Controllable abstractive dialogue summarization with sketch supervision," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5108–5122. [Online]. Available: https://aclanthology.org/2021.findings-acl.454

[23] N. Shashi *et al.*, "Planning with Learned Entity Prompts for Abstractive Summarization," *arXiv e-prints*, p. arXiv:2104.07606, Apr. 2021.

[24] J. Chen and D. Yang, "Structure-aware abstractive conversation summarization via discourse and action graphs," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 1380–1391. [Online]. Available: https://aclanthology.org/2021.naacl-main.109

[25] L. Mike *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[26] Z. Jingqing *et al.*, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 11 328–11 339. [Online]. Available: https://proceedings.mlr.press/v119/zhang20ae.html

[27] L. Dong *et al.*, "Unified language model pre-training for natural language understanding and generation," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach *et al.*, Eds., 2019, pp. 13 042–13 054. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html

[28] Y. Zhang *et al.*, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 270–278. [Online]. Available: https://aclanthology.org/2020.acl-demos.30

[29] J. Chen and D. Yang, "Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization," in *EMNLP*, 2020.

[30] H. Yichong *et al.*, "The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey," *arXiv e-prints*, p. arXiv:2104.14839, Apr. 2021.

[31] Z. Zheng *et al.*, "Reducing Quantity Hallucinations in Abstractive Summarization," *arXiv e-prints*, p. arXiv:2009.13312, Sep. 2020.

[32] C.-Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*, July 2004, pp. 74–81. [Online]. Available: https://www.microsoft.com/en-us/research/publication/rouge-a-package-for-automatic-evaluation-of-summaries/

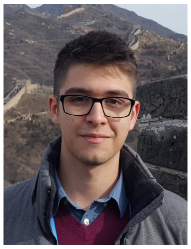
[33] C. Sammut and G. I. Webb, Eds., *TF-IDF*. Boston, MA: Springer US, 2010, pp. 986–987.

[34] S. Qaiser and R. Ali, "Text mining: Use of tf-idf to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, 07 2018.

[35] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[36] M. H. Bokaei *et al.*, "Summarizing meeting transcripts based on functional segmentation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 10, p. 1831–1841, Oct. 2016. [Online]. Available: https://doi.org/10.1109/TASLP.2016.2585859

[37] A. R. Fabbri *et al.*, "SummEval: Re-evaluating Summarization Evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 04 2021. [Online]. Available: https://doi.org/10.1162/tacl_a_00373



**George Pantelimon Prodan** received the B.S. degree in Computer Science from the Ovidius University of Constanta, Romania and B.S. degree in Physics from the University of Bucharest, Romania. He is currently pursuing his M.S. studies in Physics of Data at the Department of Physics and Astronomy, University of Padua, Italy. His research interests image and language processing and their general applications in life data and astronomy.

**Elena Pelican** is Assoc. Prof. of Optimization, Machine Learning, and Data Mining at the Department of Mathematics and Computer Science, Ovidius University of Constanta, Romania. She earns a PhD from Politehnica University of Bucharest, Romania in 2007, and she is the holder of "2009 Young Researcher of the Year" awarded by the Romanian Industrial and Applied Mathematics Society. Her current research interests are pattern recognition and data mining with applications in natural language processing and computer vision.

## APPENDIX A
## TWO SHOT PROMPT EXAMPLE

In Table VI we show one example of a two-shot training to illustrate the prompt structure for summary generation.

## APPENDIX B
## WEIGHT EXPERIMENTS

In this appendix we present the experimental results for several weight configurations (Table VII). A number of 50 tests are performed for each experiment at the temperature of 0. The average token count (ATC) and the ROUGE scores are provided.

TABLE VI
EXAMPLE OF A TWO-SHOT TRAINING

| **Item** | **Prompt** |
|---|---|
| Hint | Summarize conversations |
| Delimiter | """ |
| Shot 1 | Jackson: file-gif<br>Jackson: file-gif<br>Jackson: file-gif<br>Jackson: file-gif<br>Madison: LOL dude stop it, youre being such a little troll right now<br>Jackson: file-gif<br>Jackson: file-gif<br>Jackson: file-gif<br>Madison: STOP lol<br>Jackson: file-gif<br>Jackson: :) |
| Delimiter | """ |
| Summary 1 | Summary: Jackson is sending Madison a lot of gifs. |
| Delimiter | """ |
| Shot 2 | Paul: Hey Rosie ¡3<br>Rosemary: Hey Paolooo<br>Paul: i had so much fun last time we met<br>Paul: i miss you already<br>Rosemary: miss you too<br>Rosemary: file-gif<br>Paul: file-gif<br>Paul: file-gif<br>Paul: yesterday i quit my job<br>Rosemary: wait what? :(<br>Paul: yeah im pissed<br>Paul: they fired most of my friends<br>Paul: so i quit<br>Rosemary: :(<br>Paul: dont worry ive been saving money for long time<br>Paul: i'll find sth soon :¿<br>Rosemary: good luck :O |
| Delimiter | """ |
| Summary 2 | Summary: Paul quit his job yesterday because most of his friends got laid off. He has money saved and will find a new job soon. |
| Delimiter | """ |
| Dialogue<br>to summarize | Hannah: Hey, do you have Betty's number?<br>Amanda: Lemme check<br>Hannah: file-gif<br>Amanda: Sorry, can't find it.<br>Amanda: Ask Larry<br>Amanda: He called her last time we were at the park together<br>Hannah: I don't know him well<br>Hannah: file-gif<br>Amanda: Don't be shy, he's very nice<br>Hannah: If you say so..<br>Hannah: I'd rather you texted him<br>Amanda: Just text him :)<br>Hannah: Urgh.. Alright<br>Hannah: Bye<br>Amanda: Bye bye |
| Delimiter | """ |
| Summary<br>to complete | Summary: |

TABLE VII
EXPERIMENTAL RESULTS FOR SEVERAL WEIGHT CONFIGURATIONS

| Dataset | $w_c$ | $w_s$ | $w_a$ | Shots | ATC | R1 | R2 | RL |
|---|---|---|---|---|---|---|---|---|
| SCd | 0.6 | 0.3 | 0.1 | 3 | 309.76 | 41.72 | 15.17 | 32.13 |
| SCd | 0.1 | 0.5 | 0.4 | 2 | 258.3 | 41.28 | 16.49 | 32.31 |
| SCd | 0.1 | 0.6 | 0.3 | 2 | 258.66 | 41.2 | 16.56 | 32.37 |
| SCd | 0.1 | 0.3 | 0.6 | 2 | 255.98 | 41.04 | 15.86 | 31.78 |
| SCd | 0.1 | 0.4 | 0.5 | 2 | 257.02 | 41.01 | 15.46 | 31.57 |
| SCd | 0.2 | 0.5 | 0.3 | 2 | 256.06 | 40.63 | 16.03 | 31.9 |
| SCd | 0.2 | 0.6 | 0.2 | 2 | 255.98 | 40.61 | 15.47 | 31.31 |
| SCd | 0.1 | 0.3 | 0.6 | 3 | 317.54 | 40.44 | 14.57 | 30.79 |
| SCd | 0.5 | 0.1 | 0.4 | 2 | 235.06 | 40.41 | 14.52 | 30.63 |
| SCd | 0.1 | 0.5 | 0.4 | 3 | 319.1 | 40.38 | 14.57 | 31.33 |
| SCd | 0.3 | 0.5 | 0.2 | 3 | 314.02 | 40.35 | 14.65 | 30.5 |
| SCd | 0.3 | 0.5 | 0.2 | 2 | 254.16 | 40.27 | 15.93 | 31.55 |
| SCd | 0.6 | 0.1 | 0.3 | 2 | 229.84 | 39.86 | 14.74 | 30.65 |
| SCd | 0.3 | 0.6 | 0.1 | 2 | 255.54 | 39.69 | 15.33 | 30.42 |
| SCd | 1 | 0 | 0 | 2 | 240.84 | 39.67 | 15.15 | 32.22 |
| SCd | 0.3 | 0.1 | 0.6 | 2 | 246.38 | 39.65 | 13.32 | 29.63 |
| SCd | 0.5 | 0.3 | 0.2 | 2 | 254.54 | 39.57 | 14.11 | 30.63 |
| SCd | 0 | 0.8 | 0.2 | 2 | 260.48 | 39.51 | 12.71 | 29.47 |
| SCd | 0.1 | 0.6 | 0.3 | 3 | 320.2 | 39.45 | 13.71 | 31.31 |
| SCd | 0.6 | 0.2 | 0.2 | 2 | 244.56 | 39.39 | 13.29 | 29.62 |
| SCd | 0.3 | 0.3 | 0.4 | 2 | 253.94 | 39.35 | 14.79 | 30.45 |
| SCd | 0.3 | 0.4 | 0.3 | 2 | 255.04 | 39.34 | 14.96 | 30.77 |
| SCd | 0.1 | 0.1 | 0.8 | 2 | 253.94 | 39.2 | 14.48 | 29.75 |
| SCd | 0.2 | 0.2 | 0.6 | 2 | 253.94 | 39.09 | 14.37 | 29.72 |
| SCd | 0 | 1 | 0 | 2 | 260.36 | 38.99 | 12.6 | 29.64 |
| SCd | 0.8 | 0.2 | 0 | 2 | 241.36 | 38.94 | 13.98 | 30.13 |
| SCd | 0.5 | 0.4 | 0.1 | 2 | 253.9 | 38.92 | 14.39 | 29.8 |
| SCd | 0.5 | 0.4 | 0.1 | 2 | 253.9 | 38.92 | 14.39 | 29.8 |
| SCd | 0.8 | 0 | 0.2 | 2 | 258.26 | 38.87 | 11.97 | 29.93 |
| SCd | 0.4 | 0.3 | 0.3 | 2 | 253.96 | 38.79 | 14.35 | 29.58 |
| SCd | 0.6 | 0.3 | 0.1 | 2 | 248.28 | 38.68 | 13.99 | 30.32 |
| SCd | 0.2 | 0 | 0.8 | 2 | 244.98 | 38.45 | 11.87 | 29.96 |
| SCd | 0 | 0.2 | 0.8 | 2 | 260.48 | 38.44 | 15.46 | 29.15 |
| SCd | 0.1 | 0.6 | 0.3 | 1 | 129.64 | 38.13 | 13.36 | 29.5 |
| SCd | 0.1 | 0.5 | 0.4 | 1 | 129.48 | 38.12 | 13.6 | 29.83 |
| SCd | 0.1 | 0.3 | 0.6 | 1 | 127.98 | 38.12 | 13.38 | 29.61 |
| SCd | 0 | 0.5 | 0.5 | 2 | 260.48 | 38.08 | 12.53 | 29.16 |
| SCd | 0.2 | 0.8 | 0 | 2 | 255.8 | 37.69 | 12.64 | 28.24 |
| MSd | 0.1 | 0.6 | 0.3 | 2 | 517.26 | 37.63 | 16.52 | 28.5 |
| MSd | 0.3 | 0.5 | 0.2 | 2 | 516.16 | 37.61 | 16.73 | 28.1 |
| MSd | 0.1 | 0.5 | 0.4 | 2 | 516.5 | 37.39 | 16.47 | 28.37 |
| SCd | 0.3 | 0.5 | 0.2 | 1 | 127.64 | 37.38 | 12.42 | 29.57 |
| MSd | 0.1 | 0.4 | 0.5 | 2 | 516.02 | 37.19 | 16.12 | 27.82 |
| MSd | 0.3 | 0.4 | 0.3 | 2 | 515.56 | 37.16 | 16.57 | 28.12 |
| MSd | 0.2 | 0.6 | 0.2 | 2 | 515.94 | 36.82 | 16.35 | 27.69 |
| MSd | 0.2 | 0.5 | 0.3 | 2 | 515.16 | 36.75 | 15.92 | 27.43 |
| MSd | 0.2 | 0.8 | 0 | 2 | 515.94 | 36.58 | 15.8 | 27.44 |
| MSd | 0.1 | 0.3 | 0.6 | 2 | 515.94 | 36.35 | 15.81 | 27 |
| MSd | 0.5 | 0.3 | 0.2 | 2 | 512.7 | 36.18 | 15.75 | 27.27 |
| SCd | 0 | 0 | 1 | 2 | 289.9 | 35.75 | 12.57 | 28.21 |
| MSd | 0.4 | 0.3 | 0.3 | 2 | 515.28 | 35.66 | 15.52 | 26.98 |
| SCd | 1 | 0 | 0 | 1 | 120.42 | 35.56 | 10.59 | 27.31 |
| MSd | 0.5 | 0.5 | 0 | 2 | 515.56 | 35.5 | 15.31 | 27.1 |
| MSd | 0 | 0.5 | 0.5 | 2 | 518.84 | 33.34 | 12.9 | 24.37 |
| DSd | 0.1 | 0.6 | 0.3 | 2 | 350.72 | 32.92 | 9.85 | 25.95 |
| DSd | 0.3 | 0.4 | 0.3 | 2 | 348.46 | 32.84 | 10.03 | 25.69 |
| DSd | 0.1 | 0.3 | 0.6 | 2 | 348.98 | 32.83 | 9.91 | 25.36 |
| DSd | 0.6 | 0.2 | 0.2 | 2 | 344.08 | 32.81 | 9.86 | 25.67 |
| DSd | 0.3 | 0.4 | 0.3 | 3 | 490.68 | 32.62 | 9.29 | 25.66 |
| DSd | 0.1 | 0.5 | 0.4 | 3 | 496.52 | 32.55 | 8.69 | 25.15 |
| DSd | 0.5 | 0.3 | 0.2 | 2 | 346.66 | 32.53 | 10.38 | 25.49 |
| DSd | 0.1 | 0.5 | 0.4 | 2 | 350.56 | 32.52 | 10.01 | 25.02 |
| DSd | 0.1 | 0.3 | 0.6 | 3 | 494.38 | 32.49 | 8.67 | 25.44 |
| DSd | 0.4 | 0.3 | 0.3 | 2 | 346.86 | 32.41 | 10.02 | 25.44 |
| DSd | 0.2 | 0.5 | 0.3 | 2 | 349.02 | 32.3 | 9.34 | 25.42 |
| DSd | 0.2 | 0.8 | 0 | 2 | 350.46 | 32.23 | 9.66 | 25.28 |
| DSd | 0.1 | 0.6 | 0.3 | 3 | 496.82 | 31.88 | 8.97 | 24.83 |
| DSd | 0.1 | 0.4 | 0.5 | 2 | 350.46 | 31.62 | 9.69 | 24.62 |
| DSd | 0.8 | 0.2 | 0 | 2 | 342.74 | 31.08 | 9.03 | 24.6 |
| DSd | 0.1 | 0.3 | 0.6 | 1 | 174.8 | 30.27 | 8.93 | 24.37 |
| DSd | 0.3 | 0.4 | 0.3 | 1 | 175.32 | 30.19 | 8.58 | 24.14 |
| DSd | 0.1 | 0.6 | 0.3 | 1 | 175.54 | 29.73 | 8.45 | 23.57 |
| DSd | 0.1 | 0.5 | 0.4 | 1 | 175.82 | 29.64 | 8.78 | 23.64 |