

# ***Access Granted: Deep Learning Text-Generation Model for Scientific Grant Proposals***

**Julia Bobrovskiy**

NIKE, Inc.

`jbobbro@berkeley.edu`

**Ethan Duncan**

NASA Goddard Spaceflight Center

`ejduncan@berkeley.edu`

## **Abstract**

The process of writing and submitting scientific grant proposals is tedious. We hypothesize that there are a set of features that make one grant proposal better than another for a particular submission, hence making it more likely to be accepted. We aim to use existing state-of-the-art NLP algorithms to generate a scientific grant draft that will be successfully accepted at the National Science Foundation (NSF). We propose having a given author submit a sample first statement or thesis and the model will generate the rest of the abstract that authors can then use as a foundation for their submission. We fine-tuned a GPT-2 model using data from successful National Science Foundation abstracts that has over 100,000 instances from 1990-2003. We use a mix of transformer models to perform this task exploring a baseline model, T5, GPT-2 and fine-tuned GPT-2 for creation of a rudimentary successful abstract.

## **1 Introduction**

The original promise of computers was to solve the information overload of science. Although initial computers were specialized for retrieval and storage of information and not pattern recognition or science. As a result, we have a plethora of information at our finger tips, but not the means to process it. Scientists, like most of us, face the dilemma of prioritizing tasks in order to make the most of their time; unfortunately most of their time is eaten up by the task of writing, editing, proofreading, formatting and submitting scientific grants in order to get funding for their labs, their projects and their university.

Writing and submitting scientific grant proposals is tedious. For new research grants there is a 20% - 30% acceptance rate and, by some estimates, many top researchers spend up to 50% of their research time writing grant proposals. Grant funding is the way that a large number of researchers

in the STEM (Science, Technology, Engineering and Mathematics) field get their funding. Without money from scientific grants and other organizations, researchers would not be able to conduct research, publish relevant papers in scientific journals and importantly would not be competitive for further grants and positions. It is no exaggeration that winning grants and the right time in the career of a researcher can propel or hinder their career.

We would like to acknowledge at the beginning of this project that there is an overwhelming majority of men, especially caucasian men, in the STEM field. In future iterations of our work, we would like to develop versions of our NLP model that account for gender, racial and other biases as to not perpetuate these biases with this technology.

## **2 Background**

The field of NLP is broad, covering topics from sentiment analysis to question answering and other forms of text generation. Back in 2020, a team of researchers built a model called TLDR (Cachola et al., 2020) which does extreme summaries of scientific papers. TLDR generation involves high source compression and requires expert background knowledge and understanding of complex domain-specific language. There are also recent examples of text generation in the theater play writing business (Rosa et al., 2020). Finally the project most closely associated, the DARE (Papanikolaou and Pierleoni, 2020) which was a biomedical text generation released in 2020, fine-tuned with half a million biomedical abstracts from PubMed.

This is an active area of research as evidence by articles (i.e. <https://www.nature.com/articles/d41586-022-03479-w>) that have been published in the past few months relating to the most ground-breaking technology, attempt to solve this very issue.

Other text generation technologies include Galactica <https://galactica.org/> and most

prominently GPT-3 <https://openai.com/api/>. There is technology available where one could write a paragraph for a scientific abstract, and just click 'Rewrite' until the user finds a cleaner version using GPT-3 online on the OpenAI website.

### 3 Methods

The data used for this work is the UCI Machine Learning Repository that has a National Science Foundation Award Abstract dataset that has over 100,000 instances from 1990-2003 found here: <https://archive.ics.uci.edu/ml/datasets/NSF+Research+Award+Abstracts+1990-2003>

The baseline model we chose is a pre-trained T5 encoder-decoder model that is trained on Common Crawl's web crawl corpus. This web scraper captures all characters it encounters, and the model provides just under 223 million total parameters including shared, encoder, and decoder parameters. We decided to use this model as our baseline to generate abstracts without any training on our existing corpus of NSF abstracts. This baseline model was used as a proof-of-concept and will be a reference model for the subsequent models in this paper. Pre-trained T5 models inherently provide access to various topics available on the internet, hence, we expect some of the scientific topics to have contextual coverage.

Our second model takes the T5-base model and trains it on our corpus of NSF abstracts using transfer learning. We do this using a simple T5 model module that is built on top of PyTorch-lightning and Transformers <https://github.com/Shivanandroy/simpleT5/>, requiring training and validation data-frames as input with the option to tune batch size, epochs and precision as hyper-parameters. We split our data into a 1000 record training and 200 record test set, predicting abstracts using a max length of 900 characters corresponding to the average length of NSF abstracts. We also capped a 325 character max length for the first two sentences of the abstract that we use as the input text to prompt the model.

We have explored predicting abstract outputs using pre-trained transformers such as GPT-2 and T5, with the purpose of identifying if these pre-trained models would inherently predict a reasonable abstract.

The final model used is a fine-tuned GPT-2 on the 100,000 instance corpus of successful NSF abstracts from 1990-2003. GPT-2 has been trained

with generic text downloaded from the internet (Wikipedia, Reddit, etc.). We split our data into approximately 1000 record training and approximately 200 record test set, predicting abstracts using a variable max length of characters.

### 4 Results and Discussion

Our baseline model out-performed our trained T5 model when comparing Rouge scores, an outcome that was not initially expected. Our baseline did well at predicting the entire abstract, particularly the Rouge1 score for the model as demonstrated Table 1. This table compares the Rouge scores when evaluating the model's predictions against the entire abstract, versus the end of the abstract (which does not encompass the first two sentences). When evaluating these results further it became evident that the baseline model was not actually predicting sensible abstracts, but instead was re-iterating the first two sentences used as the prompt to the model before padding the rest of the abstract with tokens. This inherently elevated the Rouge scores as the model was able to exactly match the n-grams of the first two sentences in the prompt, however no sensible abstracts were generated. The pre-trained T5-base model inherently picks up a lot of noise from the web, hence the model produced a replica of the first two sentences padded with gibberish text. This is a side affect of C4 ignoring markup from the HTML.

In contrast, the simple T5 model was evaluated to have lower Rouge scores across all hyper-parameter configurations compared to the baseline, but, produced sensible abstracts with purposeful language and re-wording of the first two sentences, a clear outcome of training on the shuffled corpus of NSF abstracts. This model in turn generated abstracts that would have a better chance of being accepted by the NSF. We observe when looking at Table 2. that training the model using a larger batch size and a precision of 32 provided a Rouge1 score of 0.31, close to that of the baseline model however with much better grammatical value and context. Although the simple T5 model did not beat the baseline according to the metrics, there is a clear improvement in the overall abstract. One element that we were not able to capture using this simple T5 model was the output text length. Although we set the max length of the output text to 900, the model continuously predicted much shorter abstracts, motivating our next model that

| Baseline        | Rouge 1 | Rouge 2 | Rouge L |
|-----------------|---------|---------|---------|
| Entire Abstract | 0.34    | 0.22    | 0.28    |
| End of Abstract | 0.19    | 0.03    | 0.11    |

Table 1: Rouge scores for the baseline model when comparing the model’s predictions to the entire abstract versus the end of the abstract (minus the first two sentences).

| Simple T5 | Rouge 1 | Rouge 2 | Rouge L | Batch Size | Epochs | Precision |
|-----------|---------|---------|---------|------------|--------|-----------|
| Model 1   | 0.27    | 0.12    | 0.18    | 5          | 2      | 16        |
| Model 2   | 0.27    | 0.12    | 0.19    | 5          | 2      | 32        |
| Model 3   | 0.22    | 0.11    | 0.16    | 2          | 2      | 32        |
| Model 4   | 0.31    | 0.13    | 0.20    | 10         | 2      | 32        |
| Model 5   | 0.30    | 0.14    | 0.20    | 6          | 2      | 32        |
| Model 6   | 0.24    | 0.11    | 0.17    | 3          | 10     | 32        |

Table 2: Rouge scores for Simple T5 Models trained with different hyper-parameters.

| GPT-2   | Rouge 1 | Rouge 2 | Rouge L | Batch | Epochs | Length |
|---------|---------|---------|---------|-------|--------|--------|
| Model 1 | 0.36    | 0.35    | 0.36    | 5     | 4      | 1024   |
| Model 2 | 0.21    | 0.03    | 0.10    | 2     | 5      | 768    |
| Model 3 | 0.21    | 0.03    | 0.10    | 2     | 10     | 768    |
| Model 4 | 0.18    | 0.02    | 0.08    | 2     | 2      | 768    |
| Model 5 | 0.21    | 0.30    | 0.28    | 5     | 10     | 1024   |

Table 3: Rouge scores for GPT-2 Models trained with different hyper-parameters. The Length column refers to the max length hyper-parameter.

| Input Sentence  | Actual Abstract   | Model Predicted Abstract  |
|---|---|---|
| We propose to study the influence of small-displacement faults on both saturated and unsaturated water flow and chemical transport in poorly consolidated sandy sediments. Previous work has focused on the influence of faults on saturated fluid flow in rock, rather than sediments. | ... These studies cannot be directly applied to flow in sediments because the mechanics of deformation and resultant structures are different for sediments than for rocks, and cannot be applied to unsaturated flow because the distribution of permeability in a given material is different under unsaturated versus saturated conditions ... By addressing the problem with a interdisciplinary approach, using a variety of techniques at different scales, we hope to determine whether these faults are significant hydrogeologic features. | ... In this research we will examine the effect of these fault types on sedimentary behavior by examining two important processes that occur when large quantities are stored at a low pressure (e g, e b ) or high pressure (> 100M)... For example lakes provide fertile bed habitat across northern most temperate land. But even though they are largely unharmed inland inundations cause massive flooding losses. |

Table 4: Abstract examples from the fine-tuned GPT-2 model. The rest of the abstracts can be read in the Appendix

we were able to tune with greater detail.

One of the hypothetical strengths of this entire project turned out to be its greatest weakness. Even with the best fine-tuned GPT-2 model that we created, the model does not do well generating abstracts that align with the given thesis statement and especially struggles with ideas that have fewer examples in the corpus. We encountered a common phenomenon called hallucination with our text generation model. For example, we had instances where the model would put in random scientists names to give support to something that they were not an expert on.

One of the fundamental faults of this model is that fact that most scientific work being published is novel and therefore extremely difficult to predict using an algorithm. Another aspect where the model has difficulty is filling in the appropriate usage of scientific symbols. This problem would require much more training and fine-tuning than is within the scope of this project. This model's strength is filling in the "fluff" language so that the author does not have to waste time or energy filling in these details.

## 5 Future Work

Given more time, computing resources, and successful abstract data focusing on a specific field similar to what was done in (Papanikolaou and Pierleoni, 2020). We were restricted to using around 1,000 abstracts due to limited memory on the GPU's that were used in this project. Given more time, we could have been more creative with the limited memory and done better batching with the abstracts that would have let us use more of the corpus of data available to us. Also, having a wider corpus of successful abstracts outside of the UCI dataset would have expanded the knowledge and reach of the models in the project.

## 6 Data Availability

[https://github.com/ejdunk/W266\\_Final\\_Project](https://github.com/ejdunk/W266_Final_Project)

## 7 Appendix

Below are the full result of the written and predicted abstract by the GPT-2 Model:

- **Baseline Input Sentence:** In this project, the PI's team will study the circulation of the North Pacific, from the sea surface to the

bottom, and relate it to results from an earlier study of the South Pacific. These results will be compared with circulation patterns determined by the PI in the South and North Atlantic, using similar techniques

- **Baseline Actual Abstract:** In this project, the PI's team will study the circulation of the North Pacific, from the sea surface to the bottom, and relate it to results from an earlier study of the South Pacific. These results will be compared with circulation patterns determined by the PI in the South and North Atlantic, using similar techniques. These will then feed in to studies of the ocean/ocean relationship and determination of inter-basin exchange processes

- **Baseline Predicted Abstract:** in this project, the PI's team will study the circulation of the North Pacific, from the sea surface to the bottom. results will be compared with results from an earlier study of the South and North Atlantic, using similar techniques. the project is expected to be completed by the end of January, with a final report expected in mid-October, but not before November, according to [- \*\*T5 Input Sentence:\*\* This grant provides funding for building highly scalable distributed algorithms for a new class of adaptive manufacturing enterprises using distributed agent architecture over the Internet. These algorithms will be used for reconfiguring work-in-process inventory levels and production schedules to adapt to changing market demands and supply-chain conditions
  - \*\*T5 Actual Abstract:\*\* This grant provides funding for building highly scalable distributed algorithms for a new class of adaptive manufacturing enterprises using distributed agent architecture over the Internet. These algorithms will be used for reconfiguring work-in-process inventory levels and production schedules to adapt to changing market demands and supply-chain conditions. Moreover, these algorithms and the necessary information will be embedded in agents geographically distributed throughout the enterprise. The goal is to maintain responsiveness and effectiveness to enable the next generation of scalable enterprises. Scalability will](http://cnn.com/pi-projects/aspx/sci/2005//s-gragra-nrs-elw-ly-re[hi/*-...|---de,-[*-[-,-]opc]-?.e-/_-*n_dyt(m'd,...</a></li></ul></div><div data-bbox=)

be achieved through (1) distributed algorithms that are predictable and computationally efficient; (2) distributed agent architectures that support growth in size and capability; and (3) distributed clusters that provide rapid access to information using cost effective technologies. Analytical models will be developed to predict the emergent behavior of such systems and stability and ....

- **T5 Predicted Abstract:** The algorithms will be used to reconfigure inventory levels and production schedules to adapt to changing supply-chain conditions. This project will provide funding for the development of a new class of adaptive manufacturing enterprises using distributed agent architecture over the Internet.
- **GPT-2 Input Sentence:** Research on this grant aims to investigate the optical properties of rare earth ions doped in crystalline host materials using the techniques of laser spectroscopy. Two important neodymium-based single crystals: YVO<sub>4</sub>:Nd<sup>3+</sup> (1 %) and CaMgZnY<sub>2</sub>Ge<sub>3</sub>O<sub>12</sub>:Nd<sup>3+</sup> (2%) will be studied initially.
- **GPT-2 Actual Abstract:**... There have been some studies on the first material, but none on the second. In the final phase of the project, a more complex, triply-doped material: YAG: Cr<sup>3+</sup> (1%), Tm<sup>3+</sup> (5%), Ho<sup>3+</sup> (1%) will be investigated. A complete and systematic spectroscopic investigation of these materials will be performed, their optical and laser properties will be compared with those of existing laser materials, and the potential of these materials as lasers will be evaluated. The measurements will include absorption, emission, and excitation spectra and fluorescence lifetimes at different temperatures. These materials have received much attention in recent years since their emission wavelength falls in the 2 μm range, which has proven to be very attractive for a number of medical applications because of the strong water absorption at this wavelength. The 2 μm 5I<sub>7</sub> - 5I<sub>8</sub> laser transition in Ho<sup>3+</sup> is attractive because of its long upper laser level lifetime, of the order of 5 ms, permitting larger energy storage than Nd:YAG, and its eye-safe wavelength. Study of the important mechanism of energy transfer from Cr<sup>3+</sup> to Tm<sup>3+</sup> and finally to Ho<sup>3+</sup> in this crystal will provide an excellent educational experience for undergraduates involved in the project; this grant is part of NSF's Research in Undergraduate Institutions (RUI) activity. The proposed work is in an area of current

scientific interest in the field of photonics, is technologically relevant, and has excellent potential to provide a rewarding educational opportunity for the participating undergraduates who will be exposed to modern spectroscopic equipment, techniques, and theories.

- **GPT-2 Predicted Abstract:** ... by a variety of investigators in chemistry, physics and optics. The PI has also developed several new equipment for study. This award is jointly supported by an NSF/National Science Foundation Grant (NSF 40149508), the Office for Biological Sciences' Division Undergraduate Fellowship Program at Case Western Reserve University. Undergraduates from all levels are eligible to participate in both research programs; they may choose one or more undergraduate postdoctoral appointments within their first two years under CAREER FEES program. Support through this fellowship provides support over three academic years as well as travel expenses that can accrue during these seasons. To date the research activities have been carried out with regard not only to better's laboratory but also with Dr Seddon Haldane Observatory during his tenure. There are currently approximately 400 graduate students enrolled in the US Nuclear Regulatory Commission Graduate student program. In addition to the project should provide additional training opportunities so future investigators might benefit. From her participation she hopes it would contribute significantly toward understanding how molecules interact between ion channels which control cell behavior. Furthermore such studies could potentially lead to undernrg insight into fundamental aspects of many important physical processes.

## References

- Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. [TLDR: Extreme summarization of scientific documents](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.
- Yannis Papanikolaou and Andrea Pierleoni. 2020. [DARE: data augmented relation extraction with GPT-2](#). *CoRR*, abs/2004.13845.
- Rudolf Rosa, Ondrej Dusek, Tom Kocmi, David Marecek, Tomáš Musil, Patrícia Schmidtová, Dominik Jurko, Ondrej Bojar, Daniel Hrbek, David Kosták, Martina Kinská, Josef Dolezal, and Klára Vosecká.

2020. Theaitre: Artificial intelligence to write a theatre play. *CoRR*, abs/2006.14668.