Do Hybrid and Abstractive Summarization Methods Improve Performance Over Extractive Summarization Methods in Legal Documents?

Chris John, Daniel Yi, and Ryan Schaefer
University of California, Berkeley
Berkeley, CA, United States
{chrisjohn47,daniel_yi,ryan_schaefer}@berkeley.edu

Abstract

Summarizing legislative documents can lead to distinct challenges due to their complexity, length, and structured nature. While traditional extractive methods like Cluster Extraction and MMR (Maximal Marginal Relevance) have been previously utilized due to their efficiency and interpretability, newer methods in hybrid and abstractive techniques bring a newer, potentially more promising alternative. In this paper, we aim to evaluate extractive, hybrid, and abstractive summarization methods on the BillSum dataset, a compilation of U.S. Congressional and California State bills. In specific, we explored Cluster Extraction methods combined with LED (Longformer Encoder-Decoder), as well as MMR paired with PEGASUS and BigBird PEGASUS. Our results ultimately demonstrated that hybrid methods, particularly utilizing Cluster Extraction combined with LED, outperform other purely extractive or abstractive methods when evaluating both ROUGE and BERTScore metrics. This indicated improved coherence, fluency, and semantic value within our generated summaries. These findings help point out the effectiveness in implementing extractive frameworks within large summarization models for legal summarization tasks.

1 Introduction

Every year, over 10,000 bills are introduced in the US Congress, with tens of thousands more introduced in state legislatures. Summarization of these bills and additional legal documents is critical for enabling legislators, journalists, and the public to quickly understand complex legislative content, track policy changes, and make informed decisions. For research toward this subject, Kornilova and Eidelman (2019) introduced the BillSum dataset, a large-scale corpus of U.S. Congressional and California state bills paired with human-written summaries. This dataset has since become a benchmark for evaluating summarization models in the legal domain. A model that performs well in summarizing the documents of the BillSum dataset will be able to handle similarly structured legislative texts, making it a strong candidate for real-world applications in legal domains.

However, summarizing complex legal documents comes with its own set of challenges. Legislative documents are typically lengthy, formal, and structurally complex. Their use of specialized terminology, nested clauses, and cross-references adds to the challenge, especially for automated systems attempting to generate accurate summaries. Previous work in legal summarization has favored extractive models, which select and reorder key sentences or phrases directly from the source text. However, with the rise of hybrid and abstractive methods like PEGASUS and T5, there has been growing interest in models that can generate more fluent and human-like summaries. In this paper, we explore extractive, hybrid, and abstractive models for legal text summarization, examining their strengths, limitations, and suitability for handling complex legislative documents.

2 Background

As mentioned previously, extractive methods have been commonly used for summarization tasks. Algorithms such as Maximal Marginal Relevance (MMR), which have been applied to legal documents (Agarwal et al., 2022), are fully unsupervised and require relatively low computational resources. Additionally, MMR effectively reduces redundancy, which is commonly found in legal documents, by selecting sentences that provide maximum informational value while minimizing overlap.

Deep clustering, which combines sentence-level summary relevance prediction with a cluster-based extractive technique, has also been proposed for legal document summarization (Jain et al., 2024). Cluster-based methods group semantically simi-

lar sentences and select representative ones from each cluster. Often, particularly in legal texts, they consider the document structure more holistically, potentially leading to more balanced and informative summaries.

However, not all extractive methods perform equally well. Simple bag-of-words and frequency-based models often underperform relative to more advanced approaches, particularly when evaluated using ROUGE metrics (Jain et al., 2021). Motivated by these findings, we build upon a baseline of simple extractive summarization techniques by integrating additional semantic models such as PEGASUS to better capture the contextual relevance and generate more coherent summaries of legislative texts.

Looking further, we find that abstractive models can offer improvements in both readability and summarization accuracy. Zhang et al. (2020) explores the summarization performance of various text domains using pretrained models such as PEGASUS. They report significant improvements in ROUGE scores across multiple summarization tasks, including BillSum. Drawing from their results, we also incorporate PEGASUS into our pipeline to evaluate its effectiveness on legislative texts, particularly in comparison to extractive models.

3 Methods

3.1 Task

The overarching goal of our work is to develop multiple summarization models and compare their performance. We use two extractive methods—cluster-based and MMR—as our baseline and evaluate them alongside more advanced approaches, including hybrid and abstractive models. To assess summarization quality, we compare the generated summaries to the reference summaries provided in the BillSum dataset, employing standard evaluation metrics such as ROUGE, and newer metrics such as BERTScore.

3.2 Data

We use the BillSum dataset (Kornilova and Eidelman, 2019), a corpus of 23,455 U.S. Congressional and California State bills paired with a human-written summary, which is publicly accessible via Hugging Face. The bills are pre-split into train and test sets, with 18949 documents in the training set, 1237 documents in the CA test set, and 3269 documents in the US test set. The introduction date ranges 1998-2018 for congressional bills and 2015-2016 for California bills. These documents follow a consistent structural pattern, typically beginning with a numbered section title, followed by subsections labeled with lowercase letters, and further subdivided when necessary using numbers and uppercase letters, as shown in Figure 3[1].

		Min.	Max.	Avg.
Bills	**Tokens**	247	10470	1763.6
	Characters	5001	19998	10271.9
Summaries	**Tokens**	10	943	209.5
	Characters	52	4995	1105.5

Table 1: Token (PEASUS tokenizer) and character lengths of BillSum corpus

The BillSum corpus focuses on mid-length legislation, with documents ranging from 5,000 to 20,000 characters in length. Each bill is accompanied by a human-written summary constrained to a maximum of 2,000 characters. As we plan to employ abstractive summarization methods such as PEGASUS, which imposes a maximum input sequence length 4096 tokens for BigBird PEGASUS—(Zaheer et al., 2021), we conduct exploratory data analysis (EDA) to examine the token-length distribution of the input bills. Table 1 presents the average, maximum, and minimum token and character lengths across the dataset and Figure 1 shows the distribution of the tokens. Since the maximum token lengths of the input exceed PEGASUS's token limit, it may be necessary to explore input shortening techniques that preserve critical information.

We also perform EDA on the human-written summaries in the BillSum dataset to better understand the target output distribution and guide hyperparameter tuning for model training and generation. As shown in Table 1, the average token length for a bill summary is 204, which is well within the generation capacity of models like PEGASUS and suitable for our compute constraints.

Finally, we examine the average number of sentences per summary to inform our decisions regarding extractive techniques. Both cluster-based and MMR extraction methods require a hyperparameter specifying the number of sentences to include in the output. Based on our analysis, the average human-written summary in the BillSum dataset contains

[1]See Appendix Section A

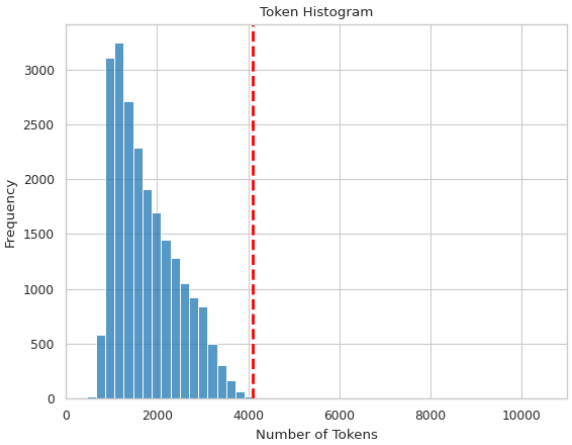


Figure 1: Distribution of token lengths of BillSum with max token indicator (BigBird PEGASUS).

approximately 5.5 sentences. Accordingly, we set the output length for these extractive approaches to 6 sentences to align closely with the reference summaries.

3.3 Evaluation Metrics

To evaluate our generated summaries in comparison to human-written ones, we use ROUGE and BERTScore.

ROUGE is widely used for evaluating text summarization systems. It is simple and easy to interpret as it measures the lexical overlap between generated and reference summaries using n-grams and longest common subsequences. However, ROUGE is heavily surface-form dependent and may penalize valid paraphrases or semantically equivalent rewordings. Furthermore, it does not explicitly account for fluency or coherence.

BERTScore is a relatively newer metric for summarization that leverages contextual embeddings from BERT. It compares the semantic similarity between candidate and reference summaries at the token embedding level. This allows it to better capture paraphrasing and synonymy, making it a good candidate to balance our evaluation metrics. However, BERTScore is computationally more intensive and may be sensitive to the choice of the underlying language model.

3.4 Baseline

We use two extractive models as baselines to compare against our more advanced summarization approaches—cluster-based and MMR. For extractive summarization, the initial step involves segmenting the text into sentences or sentence-like units. We segment the bills based on the numbered section titles, followed by subsections labeled with lowercase letters, since conventional sentence splitting proves challenging due to the structured format of legal documents.[2] This approach provides a balance between segment length and information retention.

We use Term Frequency–Inverse Document Frequency (TF-IDF) to vectorize each segmented unit for our extractive baselines. Despite its limitations in capturing semantic and contextual information (Jain et al., 2021), TF-IDF remains a common baseline due to its simplicity, interpretability, and low computational overhead.

3.4.1 Cluster Extraction

For cluster extraction, we use the sentence vectors generated by TF-IDF and perform clustering based on cosine distance. Guided by our EDA, we set the number of clusters to six, aligning with the average number of sentences in the human-written summaries. To extract a representative sentence from each cluster, we select the sentence whose vector is closest to the cluster centroid, capturing the most central content within that group.

3.4.2 MMR Extraction

Similar to cluster-based extraction, we utilize TF-IDF vectors for MMR and compute pairwise distances using cosine similarity. We designate the first sentence as the centroid for relevance scoring, motivated by the observation that the first sentence often contains the title or summary statement of the bill. As with the cluster-based extraction method, we fix the number of extracted sentences to six, aligning with the average length of human-written summaries in the BillSum dataset. The trade-off parameter λ in the MMR scoring function is set to 0.7, prioritizing relevance to the document while still penalizing redundancy.

3.5 Hybrid Models

For our hybrid models, we build upon our existing extractive frameworks. This approach is motivated by our interest in directly measuring the performance improvements gained by incorporating abstractive techniques on top of extractive baselines. Additionally, many models have token limits and are computationally intensive. Shortening the input

[2] See Appendix Section A for an example.

beforehand reduces compute time and information loss.

3.5.1 Cluster + Longform Encoder-Decoder

Our first hybrid model builds upon the key strengths on traditional extractive methods by integrating them with a few abstractive modeling techniques. The Cluster + LED model leverages the same initial extraction step, largely involving cluster-based sentence selection, to produce both concise and informative inputs tailored specifically for an abstractive summarization model like an LED.

We first applied a TF-IDF vectorization to segment legislative documents into sentence clusters, further selecting sentences that were more representative of each cluster using cosine similarity. We then fed the cluster-selected sentences into the LED model. LED in particular is well-suited for handling large inputs due to its sparse attention mechanism, efficiently identifying context and long range dependencies that may exist across large texts (perfect for summarization complex and well structured legal texts).

3.5.2 MMR + PEGASUS

Our second hybrid model builds on the output of the MMR-based extractive summarization. The selected sentences are concatenated and fed into BigBird PEGASUS. We retain the same output length of six sentences, consistent with our extractive baselines. This choice enables the model to capture the key information needed for summarization while ensuring that inputs remain below the token limit, as illustrated in Figure 2.

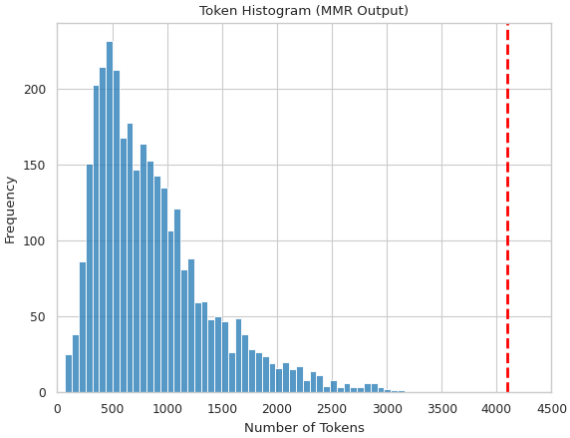


Figure 2: Distribution of token lengths of MMR output with max token indicator (BigBird PEGASUS).

3.6 BigBird PEGASUS

BigBird PEGASUS is a transformer model that uses a sparse attention approach to improve performance on long texts over full attention transformers. While PEGASUS-large has a maximum input length of 1024 tokens, BigBird PEGASUS has a maximum input length of 4096 tokens, which is better suited to our bill summarization. PEGASUS would have only been able to fully process a small portion of the documents in BillSum, but BigBird PEGASUS is able to fully process almost all of them.

Additionally, we performed fine-tuning on the model. BigBird PEGASUS was pretrained to summarize scientific papers from arxiv, which may have been slightly different to a legal document, so we evaluated the model's performance with the pretrained weights and after fine-tuning on BillSum. Due to time and computational resource constraints, we fine-tuned on a subset of 5000 bills from the training data for 1 epoch.

4 Results and Discussion

Looking at the results, we find that overall performance aligns with our expectations, with each model performing in line with its complexity and design—though a few unexpected outcomes also emerged.

4.0.1 Baseline

The Cluster and MMR extractive models had the lowest ROUGE/BERTScores and yielded nearly identical results, as shown in Table 2. We attribute this similarity to the dominant influence of sentence vectorization in these low-cost models, which likely limits the variance introduced by differing sentence selection strategies. In both models, we used TF-IDF for sentence vectorization, a frequency-based approach that is considered somewhat outdated compared to more context-aware embeddings. Using more up-to-date sentence embedding strategies, such as SentenceTransformers (Reimers and Gurevych, 2019), may result in better performance and a more nuanced comparison.

4.0.2 Hybrid Models

For our hybrid models, we find that the evaluation results from our Cluster + LED model significantly outperformed both baseline and event abstractive methods in both ROUGE and BERTScore metrics. This indicates better semantic coherence

Model	ROUGE				BERTScore		
	rouge-1	**rouge-2**	**rouge-L**	**rouge-Lsum**	**Precision**	**Recall**	**F1**
Cluster	0.3137	0.1725	0.2114	0.2772	0.7324	0.8399	0.7817
MMR	0.3117	0.1730	0.2010	0.2751	0.7367	0.8407	0.7845
Cluster + LED	**0.4376**	**0.2629**	**0.3229**	**0.3599**	**0.8931**	**0.8659**	**0.8788**
MMR + BigBird PEGASUS	0.1759	0.0242	0.1169	0.1301	0.7981	0.7877	0.7924
BigBird PEGASUS	0.1931	0.0253	0.1259	0.1419	0.7907	0.7893	0.7896
Fine-Tuned BigBird PEGASUS	0.2433	0.0409	0.1534	0.1762	0.8147	0.8004	0.8072

Table 2: ROUGE and BERTScore results for proposed models.

with human-written summaries, along with preserving semantic relevance and generating human-like summaries, as shown in an F1 score of 0.8788. These findings largely highlight the effectiveness of hybrid summarization methods, particularly combining cluster-based extraction methodologies with perfect long-document parsers that the LED model can parse through. This combination addresses a lot of the unique challenges posed by long-form legal document summarization.

On the other hand, the MMR + BigBird PEGASUS model underperformed, achieving only a marginally higher F1 BERTScore compared to the extractive baseline (0.7924). We attribute this to the fixed output length of six sentences, which likely constrained the model's ability to fully leverage its abstractive capabilities and synthesize broader contextual information from the input. However, it is important to note that the Cluster + LED model also operated under the same six-sentence constraint, yet still achieved noticeably better performance. This suggests that LED may be more efficient at leveraging limited input to generate coherent and informative summaries. Further experimentation could be helpful in determining the exact reasons for this difference.

4.1 Abstractive Models

The evaluation results for BigBird PEGASUS without fine-tuning show that the model performed very similarly the MMR + BigBird PEGASUS hybrid approach. This indicates that extractive approaches can provide an adequate approximation of the full text for the model to generate summaries of similar quality. The benefit of this is that the MMR text is significantly shorter than the full text, meaning the model is less resource intensive in generating summaries and can run faster.

Fine-tuning BigBird PEGASUS did noticeably improve all of the recorded metrics. As mentioned in the methods section, we only fine-tuned on a subset of 5000 records from the training data, which is about a quarter of the available data, for 1 epoch. Further research is required to determine how much performance would improve using the full training set and to determine the optimal number of training epochs.

Both the base BigBird PEGASUS model and the fine-tuned model outperformed the baseline in precision and F1 BERTScore. However, both models performed worse than the baseline in recall BERTScore and all of the ROUGE metrics. This underperformance can be explained by how abstractive models generate summaries. Abstractive models do not copy text directly from the source like extractive models. Instead, they generate novel text using the source as a reference. This often results in summaries that are syntactically similar to the reference summaries, but may use different words to convey the same meaning. This results in a penalty on surface-form metrics like ROUGE. Similarly, BERTScore recall compares the embedding of each token in the reference to its most similar token in the candidate. The penalty BERTScore recall puts on abstractive generation is less severe than the penalty put on it by ROUGE metrics, but it can still be significant if the model significantly paraphrased the reference summary. BERTScore recall can also perform worse if the generated summaries are too short and do not cover everything the reference summaries cover.

5 Conclusion

We applied several extractive, abstractive, and hybrid approaches to generating summaries on the BillSum dataset. The results show that the hybrid and abstractive models can outperform the extractive models in semantics-based metrics like BERTScore, but can struggle with more surface-form metrics like ROUGE. We found that extrac-

tive methods can be used to create an approximation of the full text with a lower input token count, reducing the time and computational resources required to fine-tune and generate predictions with abstractive models without significantly impacting the quality of the generated summaries. Clustering in combination with LED in particular proved to be very effective at this task. Further research can test more combinations of extractive and abstractive models to create the best hybrid approach for long-form text.

6 Authors Contribution's

Below we list the authors contributions for this paper (all authors contributed to the Methods/Results and Discussion sections):

Daniel:
- Extractive Models (MMR, Cluster)
- MMR + PEGASUS Hybrid Model
- Introduction, Background, Slides (Introduction, Extractive)

Chris:
- Hybrid Models (LED, LED + Cluster, Hypertuning)
- LED + Cluster Hybrid Model
- Abstract, Slides (Hybrid, Results)

Ryan:
- Abstractive Models (BigBird PEGASUS)
- Conclusion, Slides (Abstractive, Conclusion)

References

Abhishek Agarwal, Shanshan Xu, and Matthias Grabmair. 2022. Extractive summarization of legal decisions using multi-task learning and maximal marginal relevance. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1857–1872, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Automatic summarization of legal bills: A comparative analysis of classical extractive approaches. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 394–400.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2024. A sentence is known by the company it keeps: Improving legal document summarization using deep clustering. *Artif. Intell. Law*, 32(1):165–200.

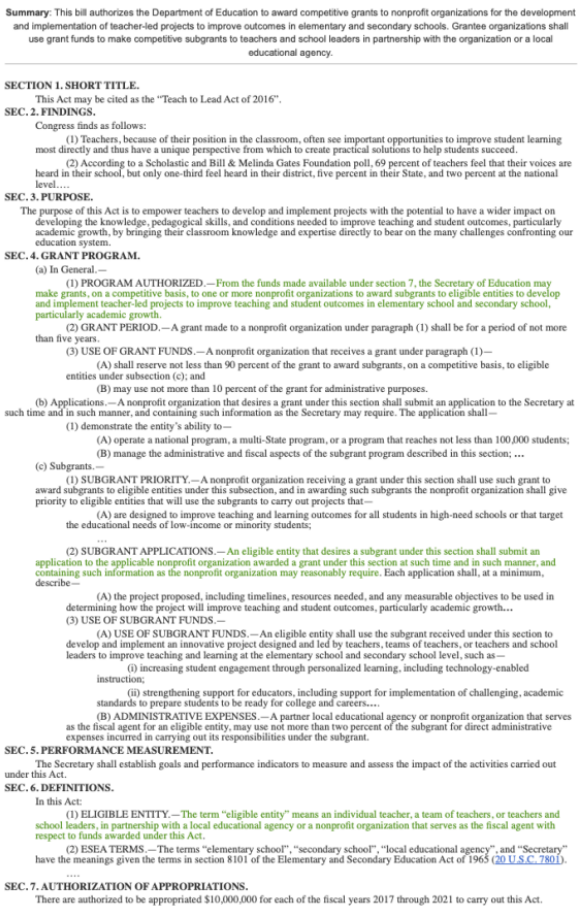
Anastassia Kornilova and Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, page 48–56. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

A BillSum Bill Structure



Summary: This bill authorizes the Department of Education to award competitive grants to nonprofit organizations for the development and implementation of teacher-led projects to improve outcomes in elementary and secondary schools. Grantee organizations shall use grant funds to make competitive subgrants to teachers and school leaders in partnership with the organization or a local educational agency.

SECTION 1. SHORT TITLE.
This Act may be cited as the "Teach to Lead Act of 2016".
SEC. 2. FINDINGS.
Congress finds as follows:
(1) Teachers, because of their position in the classroom, often see important opportunities to improve student learning most directly and thus have a unique perspective from which to create practical solutions to help students succeed.
(2) According to a Scholastic and Bill & Melinda Gates Foundation poll, 69 percent of teachers feel that their voices are heard in their school, but only one-third feel heard in their district, five percent in their State, and two percent at the national level....
SEC. 3. PURPOSE.
The purpose of this Act is to empower teachers to develop and implement projects with the potential to have a wider impact on developing the knowledge, pedagogical skills, and conditions needed to improve teaching and student outcomes, particularly academic growth, by bringing their classroom knowledge and expertise directly to bear on the many challenges confronting our education system.
SEC. 4. GRANT PROGRAM.
(a) In General.—
(1) PROGRAM AUTHORIZED.—From the funds made available under section 7, the Secretary of Education may make grants, on a competitive basis, to one or more nonprofit organizations to award subgrants to eligible entities to develop and implement teacher-led projects to improve teaching and student outcomes in elementary school and secondary school, particularly academic growth.
(2) GRANT PERIOD.—A grant made to a nonprofit organization under paragraph (1) shall be for a period of not more than five years.
(3) USE OF GRANT FUNDS.—A nonprofit organization that receives a grant under paragraph (1)—
(A) shall reserve not less than 90 percent of the grant to award subgrants, on a competitive basis, to eligible entities under subsection (c); and
(B) may use not more than 10 percent of the grant for administrative purposes.
(b) Applications.—A nonprofit organization that desires a grant under this section shall submit an application to the Secretary at such time and in such manner, and containing such information as the Secretary may require. The application shall—
(1) demonstrate the entity's ability to—
(A) operate a national program, a multi-State program, or a program that reaches not less than 100,000 students;
(B) manage the administrative and fiscal aspects of the subgrant program described in this section; ...
(c) Subgrants.—
(1) SUBGRANT PRIORITY.—A nonprofit organization receiving a grant under this section shall use such grant to award subgrants to eligible entities under this subsection, and in awarding such subgrants the nonprofit organization shall give priority to eligible entities that will use the subgrants to carry out projects that—
(A) are designed to improve teaching and learning outcomes for all students in high-need schools or that target the educational needs of low-income or minority students;
...
(2) SUBGRANT APPLICATIONS.—An eligible entity that desires a subgrant under this section shall submit an application to the applicable nonprofit organization awarded a grant under this section at such time and in such manner, and containing such information as the nonprofit organization may reasonably require. Each application shall, at a minimum, describe—
(A) the project proposed, including timelines, resources needed, and any measurable objectives to be used in determining how the project will improve teaching and student outcomes, particularly academic growth...
(3) USE OF SUBGRANT FUNDS.—
(A) USE OF SUBGRANT FUNDS.—An eligible entity shall use the subgrant received under this section to develop and implement an innovative project designed and led by teachers, teams of teachers, or teachers and school leaders to improve teaching and learning at the elementary school and secondary school level, such as—
(i) increasing student engagement through personalized learning, including technology-enabled instruction;
(ii) strengthening support for educators, including support for implementation of challenging, academic standards to prepare students to be ready for college and careers....
(B) ADMINISTRATIVE EXPENSES.—A partner local educational agency or nonprofit organization that serves as the fiscal agent for an eligible entity, may use not more than two percent of the subgrant for direct administrative expenses incurred in carrying out its responsibilities under the subgrant.
SEC. 5. PERFORMANCE MEASUREMENT.
The Secretary shall establish goals and performance indicators to measure and assess the impact of the activities carried out under this Act.
SEC. 6. DEFINITIONS.
In this Act:
(1) ELIGIBLE ENTITY.—The term "eligible entity" means an individual teacher, a team of teachers, or teachers and school leaders, in partnership with a local educational agency or a nonprofit organization that serves as the fiscal agent with respect to funds awarded under this Act.
(2) ESEA TERMS.—The terms "elementary school", "secondary school", "local educational agency", and "Secretary" have the meanings given the terms in section 8101 of the Elementary and Secondary Education Act of 1965 (20 U.S.C. 7801).
...
SEC. 7. AUTHORIZATION OF APPROPRIATIONS.
There are authorized to be appropriated $10,000,000 for each of the fiscal years 2017 through 2021 to carry out this Act.

Figure 3: An example Congressional bill, from the BillSum paper.