CONSISTENT AND EFFICIENT LONG DOCUMENT UNDERSTANDING

BY

QI ZENG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Doctoral Committee:

Professor Heng Ji, Chair
Associate Professor Hanghang Tong
Assistant Professor Han Zhao
Associate Professor Lu Wang, University of Michigan
Assistant Professor Lei Li, Carnegie Mellon University

ABSTRACT

In the age of information overload, people's information needs from long documents are rapidly emerging, while people's patience for careful reading and reasoning is gradually vanishing. While people are inundated with large amounts of long textual documents covering topics in various domains, such as news, healthcare, legal service, and finance, they struggle to gain quick, concise, and accurate insights from these long and tedious documents.

The development of automatic document understanding systems promises the possibility of assisting humans in gaining insights from those long documents. Automatic systems capture and analyze the information contained in a collection of news and scientific reports in a concise and machine-understandable way. Automatic systems parse unstructured text by identifying the relations between events and entities from long complex reading for structured data usage. Automatic systems provide reliable digests by factually and consistently summarizing recent papers, reports, news, and reviews.

However, automatically understanding long documents remains a challenge because recent state-of-the-art document understanding systems are mostly built upon transformer structures and are mostly motivated, designed, implemented, and evaluated under the short-input setting. To adapt those short-input systems to long sequences, documents have to be truncated, chunked using a sliding window, or processed in parallel on multiple machines. These additional operations usually cause the loss of long-range interdependency and introduce additional costs. Therefore, this thesis focuses on developing principled and scalable methods for more consistent and efficient long document understanding. In particular, we investigate four research problems from the perspectives of consistency and efficiency:

1) Consistent Meta-review Generation. Current work on Opinion Summarization extracts and selects representing opinions on aspects of interest under the assumption that input opinions are non-controversial. Opinions in the scientific domain can be divergent, leading to controversy or consensus among reviewers, while the scientific meta-review should be consistent with the synthesized opinions from individual reviews. Therefore, we propose to benchmark scientific opinion summarization by collecting paper meta-reviews from OpenReview, proposing a Checklist-guided Iterative Introspection approach, and constructing a comprehensive evaluation framework.

2) Consistent Document Summarization. Current abstractive summarization models often generate inconsistent content, i.e. texts that are not directly inferable from the source document, are not consistent with respect to world knowledge, or are self-contradictory.

To improve the general consistency we introduce EnergySum, where we apply the Residual Energy-based Model by designing energy scorers that reflect each type of consistency and incorporating them into the sampling process.

3) Consistent Document-level Event Argument Extraction. Recent work on document-level event argument extraction models each individual event in isolation and therefore causes inconsistency among extracted arguments across events, which will further cause discrepancies for downstream applications. To address this problem, we formulate event argument consistency as the constraints from event-event relations under the document-level setting and further introduce the Event-Aware Argument Extraction (EA^2E) model with augmented context for training and inference.

4) Efficient Document Processing. Transformer-based models are inefficient in processing long sequences due to the quadratic space and time complexity in the self-attention modules. To address this limitation, we introduce two methods for self-attention acceleration, a modified Nyström method (Skyformer) to accelerate kernelized attention and stabilize training and a Sketching-based method (Skeinformer) that applies sub-sampling sketching.

ACKNOWLEDGMENTS

My greatest thanks go to my advisor Professor Heng Ji. She is the best advisor that anyone could wish for.

I would also like to express my gratitude to Professor Hanghang Tong, Professor Han Zhao, Professor Lei Li, and Professor Lu Wang, for serving on my dissertation committee and providing me with helpful feedback.

It is a great honor for me to be a member of the Blender lab, and I would like to express my gratitude to my peers in this big family, including Manling Li, Xiaodan Hu, Qiusi Zhan, Mankareet Sidhu, Hou Pong Chan, Qingyun Wang, Xiaomeng Jin, Weijiang Li, Tuan Manh Lai, Yi Fung, Charles Yu, Pengfei Yu, Zixuan Zhang, Ziqi Wang, Chi Han, Revanth Gangi Reddy, Xingyao Wang, Zhenhailong Wang, Xueqing Wu, Xiaoman Pan, Ying Lin, and Lifu Huang.

Life outside the lab was made enjoyable due to the support from my dear friends, Xigao Li, Wei Su, Zoey Sha Li, Yanru Qu, Yuan Gao, Yu Zhang, and Peiqi Huang. Special thanks to Sudo, Sumo, and MM, for their purring love.

Finally, I owe a debt of gratitude to my family, Yinglun Xu, Peanut Butter, and Honey Butter. Their constant love and support have guided me to where I stand today.

The research of this thesis was supported in part by U.S. DARPA KAIROS Program No. FA8750-19-2-1004 and U.S. DARPA AIDA Program No. FA8750-18-2-0014.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 MOTIVATIONS

Exploded and overloaded information poses challenges to people's growing information needs: How can humans read, summarize, and analyze a large amount of long financial, healthcare, scientific, or legal documents in a relatively short time? Can machines assist us? Automatic document understanding systems promise the possibility by shedding light on their ability to efficiently represent, extract, and summarize information in documents of interest.

However, understanding long documents remains a challenge because recent document understanding systems are mostly motivated, designed, implemented, and evaluated under the short-input setting. To adapt those short-input systems to long sequences, documents have to be truncated, chunked using a sliding window or processed in parallel on multiple machines. These additional operations usually cause the loss of long-range dependency and introduce additional costs.

Recently document-level setting has been widely explored in various language understanding areas, such as Neural Machine Translation [1, 2, 3, 4], Document Summarization [5, 6], Question Answering [7], and Relation Extraction [8, 9, 10, 11, 12]. These document-level tasks share the key challenges for long-sequence text processing, including the computation complexity (for example, quadratic space and time complexity in self-attention [13, 14, 15]), the lack of available datasets due to costly human annotation [16, 17], the difficulty of information aggregation across long context, and the inter-dependency among processing units.

This thesis focuses on developing principled and scalable methods for more consistent and efficient long document understanding. Long document understanding involves the process of quickly processing the given input, accurately extracting the required information from unstructured text, and presenting the extracted information for further in-depth analysis, as shown in Figure 1.1. In particular, we investigate four research problems from the perspectives of consistency and efficiency:

- **Consistent meta-review generation**: Are the collected opinions in the generated meta-reviews consistent with the comments from individual reviews and the final decision?
- **Consistent document summarization**: Are the summaries faithfully reflected the key ideas in the given documents, factual to common knowledge, and not self-contradictory?

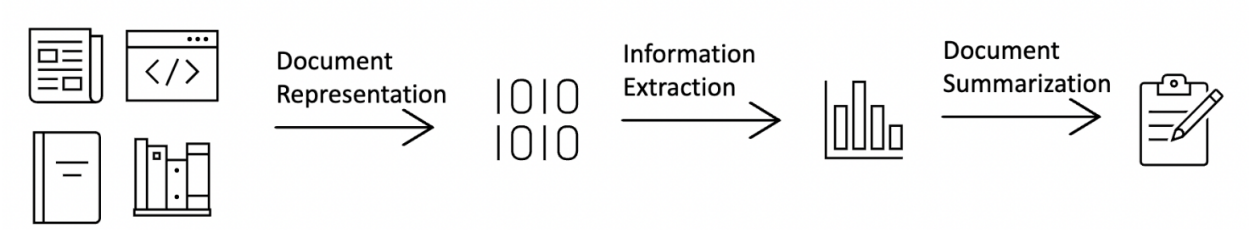

Figure 1.1: Long document understanding involves the process of document representation, information extraction, and document summarization.

Task	Type of Consistency	Object 1	Object 2
Meta-review Generation	Faithfulness	Meta-review	Individual reviews
	Self-supportiveness	Meta-review	Decisions
Document Summarization	Faithfulness	Summary	Document
	Factuality	Summary	Common Knowledge
	Self-supportiveness	Summary	Summary
Event Argument Extraction	Self-supportiveness	Event	Event

Table 1.1: Consistency means subjecting to an underlying relation pattern between two objects. We show the consistency objects and types of consistency in each task in this thesis. The taxonomy of consistency is introduced in Chapter 4.1 for document summarization and can be extended to event argument summarization and meta-review generation as well.

- **Consistent event argument extraction**: How do automatic information extractors represent and identify the interrelated event argument roles within a long document?
- **Efficient long sequence processing**: How does a document understanding system efficiently turn documents into rich representations with a limited budget of computational resources?

1.1.1 Consistency

Consistency means agreement or harmony of parts or features to one another or a whole [1]. In natural language processing, we refer to consistency as an underlying relationship pattern between two objects. The goal of improving consistency is to maintain the underlying relationship pattern. We investigate three typical language understanding scenarios with concerns about consistency, as shown in Table 1.1.

Consistency in Meta-review Generation. When the input scientific reviews from individual reviewers have different opinions, the long document understanding systems are

[1] https://www.merriam-webster.com/dictionary/consistency

required to critically summarize the controversy and consensus based on the extracted independent opinions and make decisions based on the synthesized opinions. The generated meta-review should be consistent in two ways: (1) the meta-review should reflect the opinion discussion in individual reviews, and (2) the opinion synthesis in the meta-review should be consistent with its final decision.

Consistency in Document Summarization. Current abstractive summarization models often generate inconsistent content, i.e. texts that are not directly inferable from the source document, are not consistent with respect to world knowledge, or are self-contradictory. These inconsistencies motivate a new consistency taxonomy that we define as faithfulness, factuality, and self-supportiveness. Addressing inconsistency solely in terms of faithfulness is inadequate because abstractive summarization introduces new content into the summary that is not directly copied from the source document and is not necessarily irrelevant.

Consistency in Event Argument Extraction. When the input length goes beyond the boundary of sentences, the long document understanding systems tend to lose track of the interrelations or details of the processing units in the input documents. In document-level event argument extraction, multiple events in one document are usually interconnected, and thus independent decoding method in recent work will cause contradiction among extracted arguments across events. Therefore, we study the methodologies for improving consistency in Document-level Event Argument Extraction.

1.1.2 Efficiency

Transformer-based models are not efficient in processing long sequences due to the quadratic space and time complexity of the self-attention modules, which we refer to as computation efficiency. Recent methods have been proposed to accelerate self-attention computation by selectively attending to a subset of the tokens or with low-rank matrix approximation. To avoid unreliable assumptions on the pattern selections for general usage, we instead turn to efficient approximation methods to encode documents with long transformers as a task-agnostic and scalable solution.

1.1.3 Accuracy

The ultimate goal of improving consistency and efficiency is to improve accuracy with fairly low overhead. The relations between accuracy, consistency, and efficiency are not necessarily complementary. Improvements in efficiency may come at the cost of accuracy, such as the omission of some low-level details in long document summarization, while consistency

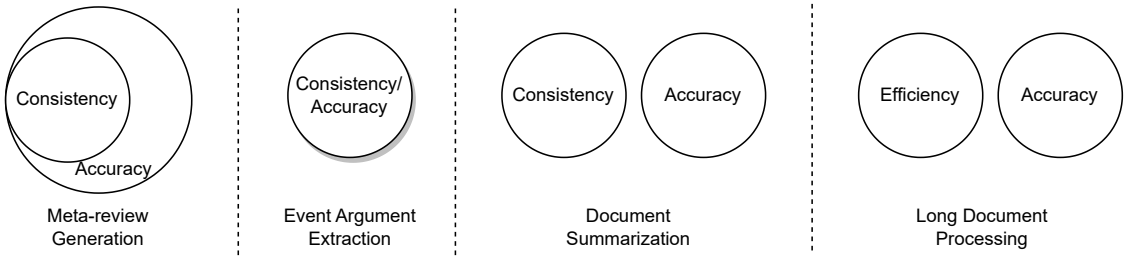

Figure 1.2: Relations of Accuracy, Consistency, and Efficiency evaluation metrics in different tasks.

improvement is usually accompanied by a heavier model design and higher computation costs.

Consistency is part of Accuracy though sometimes characterized and evaluated independently. In Consistent Meta-review Generation, discussion involvement, opinion faithfulness, and decision consistency are important evaluation dimensions. In Consistent Event Argument Extraction, improving the consistency inherently improves the accuracy because decoded inconsistent cases are considered errors in terms of accuracy evaluation. In Consistent Document Summarization, consistency is independently evaluated in addition to the overlap-based accuracy metrics though it is de facto an aspect of accuracy.

Efficiency is usually considered and evaluated independently of accuracy. In Efficient Long Sequence Processing, due to the approximation nature of the proposed methods, we focus on whether they are sufficient in getting comparable or even better accuracy than the original method while requiring fewer computation resources.

1.2 NOVELTY AND CONTRIBUTION CLAIMS

The main contribution of this thesis is to investigate the main limitations of current approaches for long document understanding and propose several new methods to solve the problems.

We first make the following assumptions to narrow down the scope of the problems and guide our methodology development:

- In consistent meta-review generation, we assume the collected meta-reviews from domain experts (Area Chairs) are inherently consistent with their underlying sentiments.
- In consistent document summarization, our assumption for the dataset is that the augmentation for the purpose of contrastive learning has captured the characteristics of all kinds of inconsistency and such feature is distinguishable and learnable.

- In consistent document-level event argument extraction, we assume the events in the document are inherently correlated and independent modeling will indirectly hurt performance.
- In efficient long sequence processing, our matrix-approximation-based methods work under the assumption that the better the modified self-attention approximates the vanilla self-attention, the better empirical performance will they have.

Here we briefly summarize the main technical contributions:

- We benchmark scientific opinion summarization by collecting a paper meta-review dataset from OpenReview, constructing a comprehensive evaluation framework, and proposing a checklist-guided iterative prompting method [18].
- We propose a consistent Residual Energy-based framework with consistency-specific energy functions and joint inference to improve consistency in document summarization [19].
- We introduce the concept of event awareness and propose an Event-Aware Argument Extraction (EA^2E) model incorporating alignment-enhanced training and iterative inference to improve consistency in event argument extraction [20].
- We introduce Skyformer, which replaces the softmax structure with a Gaussian kernel to stabilize the model training and adapts the Nyström method to accelerate the computation [14].
- We propose Skeinformer to accelerate the training and inference of transformers with initial column sampling, adaptive row normalization, and pilot sampling re-utilization [21].

1.3 THESIS OUTLINE

The rest of this proposal is structured as follows:

- Chapter 2 gives a literature survey of long document understanding tasks, including long document representation, document-level event argument extraction, and document summarization.
- Chapter 3 describes our work about benchmarking long scientific opinion summarization by building a new meta-review dataset, characterizing evaluation metrics, proposing a prompting-based method, and experimenting with state-of-the-art baseline methods.
- Chapter 4 details our energy-based method for improving consistency for document summarization.

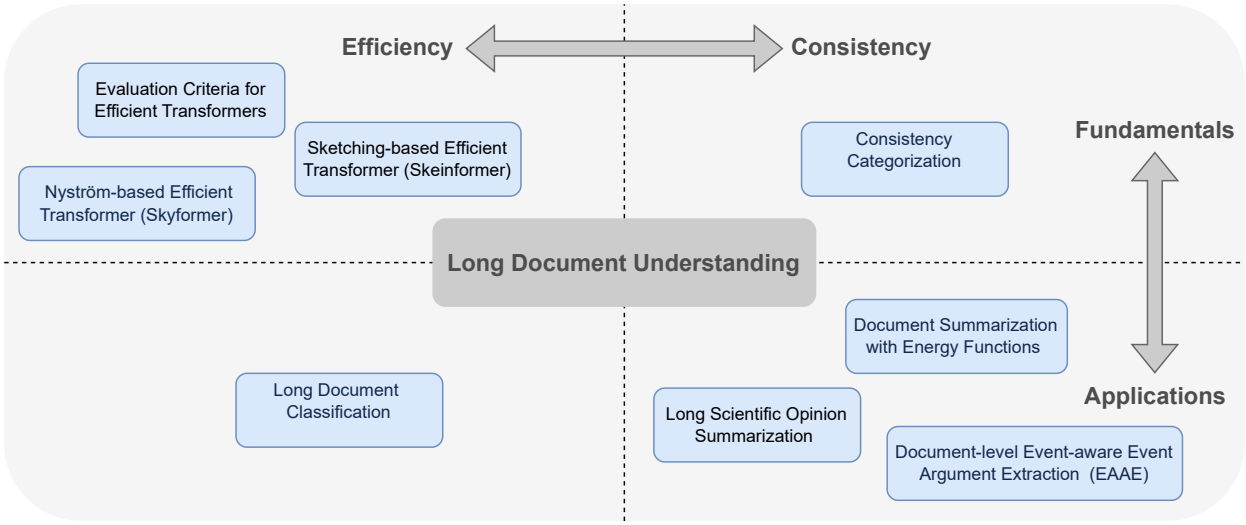

Figure 1.3: Overview of research work in this thesis.

- Chapter 5 presents our work on improving consistency for document-level event argument extraction under the assumption of a participant tends to play consistent roles across multiple events in the same document.
- Chapter 6 adapts the Nyström method to Kernelized Self-Attention in Skyformer to improve stability and efficiency in long sequence processing.
- Chapter 7 introduces Skeinformer, which applies sub-sampling sketching to reduce time complexity in long sequence processing.
- Chapter 8 concludes this thesis and discusses potential future work.

1.4 RELATED PUBLICATIONS AND IMPACT

The core of the thesis focuses on developing consistent and efficient methods for long document understanding and has a broad impact on a variety of applications: Information Extraction, Document Summarization, and many other document understanding tasks.

Additionally, some of the publications within this thesis proposal, such as Skyformer [14] and EA^{2}E [20], have technically inspired other independent follow-up research in related areas. [22] recognizes our Skyformer method as one of "the current state-of-the-art in fast self-attention operations". Transnormer [23], Fast-FNet [24] and Toeplitz Neural Network [25] follow our experimental configurations and use our method as a major baseline. CAB [26] uses the adaptive factor for ProbSparse introduced in our code repository. A survey paper for attention mechanism [27] lists our work as an important related work: "Skyformer replaces softmax with a Gaussian kernel and adapts Nyström method". SPEAE [28], SCPRG [29]

and TARA [30] use our EA^{2}E method as an important baseline for the task of document-level event argument extraction.

This thesis only includes works for which this author was the, or one of the, primary contributors. Below is the excluded research work:

- *Cross-media Structured Common Space for Multimedia Event Extraction* [31]. Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, Shih-Fu Chang. The 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020).
- *Connecting the Dots: Event Graph Schema Induction with Path Language Modeling* [32]. Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, Clare Voss. The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020).
- *ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis* [33]. Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, Nazneen Fatema Rajani. Proceedings of the 13th International Conference on Natural Language Generation (INLG 2020).
- *GAIA at SM-KBP 2020 - A Dockerlized Multi-media Multi-lingual Knowledge Extraction, Clustering, Temporal Tracking and Hypothesis Generation System* [34]. Manling Li, Ying Lin, Tuan Manh Lai, Xiaoman Pan, Haoyang Wen, Lifu Huang, Zhenhailong Wang, Pengfei Yu, Di Lu, Qingyun Wang, Haoran Zhang, Qi Zeng, Chi Han, Zixuan Zhang, Yujia Qin, Xiaodan Hu, Nikolaus Parulian, Daniel Campos, Heng Ji, Alireza Zareian, Hassan Akbari, Brian Chen, Bo Wu, Emily Allaway, Shih-Fu Chang, Kathleen McKeown, Yixiang Yao, Jennifer Chen, Eric Berquist, Kexuan Sun, Xujun Peng, Ryan Gabbard, Marjorie Freedman, Pedro Szekely, T.K. Satish Kumar, Arka Sadhu, Haidong Zhu, Ram Nevatia, Miguel Rodriguez, Yifan Wang, Yang Bai, Ali Sadeghian, Daisy Zhe Wang. Thirteenth Text Analysis Conference (TAC 2020).
- *GENE: Global Event Network Embedding* [35]. Qi Zeng, Manling Li, Tuan Lai, Heng Ji, Mohit Bansal, Hanghang Tong. Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15) at 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2021).
- *RESIN-11: Schema-guided Event Prediction for 11 Newsworthy Scenarios* [36]. Xinya Du, Zixuan Zhang, Sha Li, Pengfei Yu, Hongwei Wang, Tuan Lai, Xudong Lin, Ziqi Wang, Iris Liu, Ben Zhou, Haoyang Wen, Manling Li, Darryl Hannan, Jie Lei, Hyounghun Kim, Rotem Dror, Haoyu Wang, Michael Regan, Qi Zeng, QING LYU, Charles Yu, Carl Edwards, Xiaomeng Jin, Yizhu Jiao, Ghazaleh Kazeminejad, Zhen-

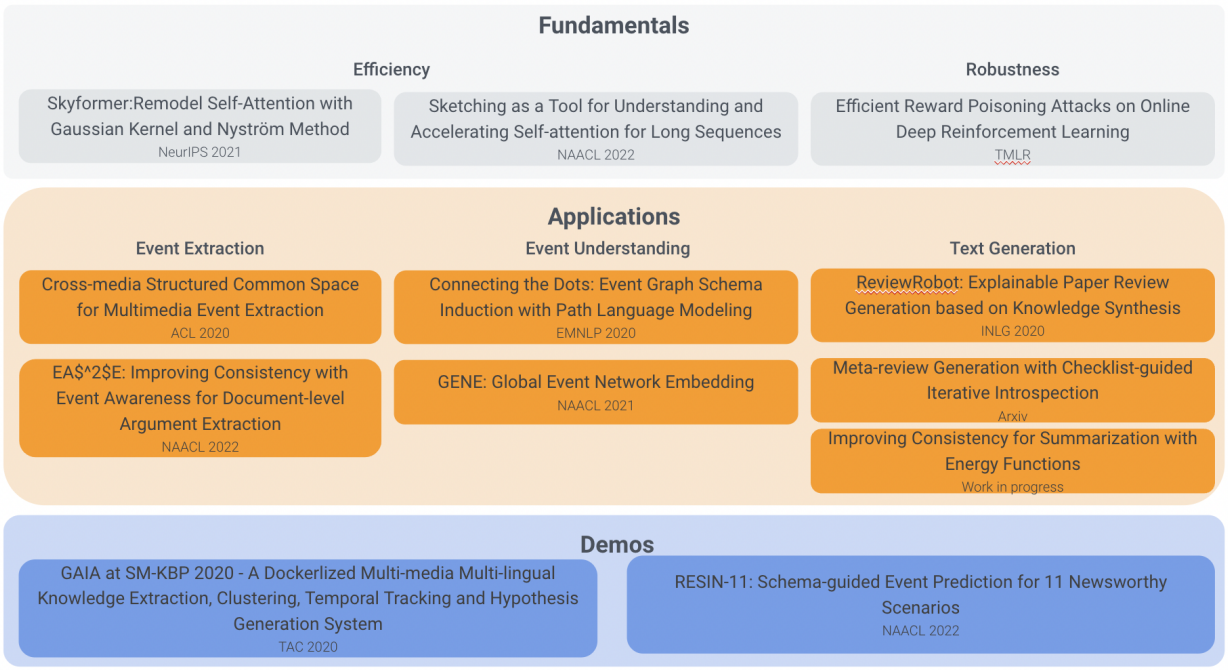

Figure 1.4: Overview of publications.

hailong Wang, Chris Callison-Burch, Mohit Bansal, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, Martha Palmer, Heng Ji. Demo track of 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022).

- *C-PMI: Conditional Pointwise Mutual Information for Turn-level Dialogue Evaluation* [37]. Liliang Ren, Mankeerat Sidhu, Qi Zeng, Revanth Gangi Reddy, Heng Ji, Chengxiang Zhai. Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering at The 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023).
- *Interpretable Automatic Fine-grained Inconsistency Detection in Text Summarization* [38]. Hou Pong Chan, Qi Zeng, Heng Ji. Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023).
- *SmartBook: AI-Assisted Situation Report Generation* [39]. Revanth Gangi Reddy, Yi R. Fung, Qi Zeng, Manling Li, Ziqi Wang, Paul Sullivan, Heng Ji.
- *Efficient Reward Poisoning Attacks on Online Deep Reinforcement Learning* [40]. Yinglun Xu, Qi Zeng, Gagandeep Singh. Transactions on Machine Learning Research (TMLR).

CHAPTER 2: LITERATURE REVIEW

To take a further step towards consistent and efficient long document understanding, in this chapter we survey the current research progress on opinion summarization, document summarization, document-level event argument extraction, and long sequence processing.

2.1 OPINION SUMMARIZATION

The task of opinion summarization is typically decomposed into aspect extraction, polarity identification, and summary generation [41]. The lack of parallel data in review opinion summaries limits the scope of most methods into the unsupervised extractive setting[42], where the aspects and sentiments from the input reviews are collected, selected, and rearranged into the output meta-reviews.

Pretrained Aspect-based Sentiment Analysis [43], variational autoencoder [44, 45], and large language models [46] enable unsupervised abstractive approaches, where the generated summaries are validated to be more fluent, informative, coherent, and concise.

To support the training and evaluation of supervised methods, recent work constructs synthetic datasets by random sampling, adding noise to the sampled summary to generate documents [47], or sampling from a Dirichlet distribution parametrized by a content planner [48]. However, the synthetic pseudo-summaries are known to be detached from real-world distributions, possibly irrelevant or inconsistent with input documents, and ignore salient latent aspects.

2.2 DOCUMENT SUMMARIZATION

Once the important information in the document has been extracted, the document understanding system is able to aggregate and rephrase the information in the form of a summary.

2.2.1 Evaluating Consistency

ROUGE score [49] measures the n-gram overlap between the generated summary and reference summary. BERTScore [50] computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings. Though effective, they do not correlate with faithful and factual correctness for generated summaries.

Automatic consistency evaluation models can be roughly classified into QA-based and entailment-based methods. QA-based methods do not require consistency labels for training. The underlying assumption for these methods is that if we ask the same questions about a summary and its source document, we will receive similar answers if the summary is factually consistent with the source. [51] introduced an IE-based method to extract a complete set of facts from both the source text and the generated summary. FEQA [52] is an automatic question-answering-based metric for faithfulness evaluation. The model first generates question-answer pairs as ground truth from the summary, then queries the document to extract answers. By comparing the answers from the summaries and from the documents, the model is able to justify whether the summary is factually consistent with respect to the questions. QAGS [53] is another automatic QA-based metric. Similar to FEQA, it measures and improves the factual consistency of summaries by asking and answering questions based on generated summaries and input documents. Questeval [54] unifies the above precision- and recall-based QA metrics and improves the correlation with human judgments over consistency, coherence, fluency, and relevance. [55] evaluates effective faithfulness of summarization systems with a faithfulness-abstractiveness trade-off curve.

Recent work has studied the use of textual entailment systems to identify factual errors, which we conclude as entailment-based methods. DAE [56] measures factual consistency by checking whether the semantic relationship manifested by individual dependency arcs in the generated summary is supported by the source document. However, the DAE score is designed to measure factuality for single sentence source sentences, and thus is not suited for longer source sequences. [57] extends DAE with generation-based and entity-based synthetic datasets for document summarization.

Due to the lack of enough costly and reliable annotation, some methods use synthetic data to train the evaluation models. [58] proposed an NLI-based fact-checking model, FactCC, which evaluates faithfulness between source documents and generated summaries with artificially corrupted datasets. [59] proposes to generate factually inconsistent summaries with masked language models as negative samples for training consistency classifiers.

2.2.2 Improving Consistency

Recent work has been looking into detecting and reducing entity-based hallucinations. [60] reduces hallucinations by integrating a language understanding module for data refinement with self-training iterations to effectively induce strong equivalence between the input data and the paired text. [61] reduces quantity hallucination by verifying quantity entities and up-ranking less hallucinated summaries. [62] proposes a loss truncation training algorithm that

filters out noisy training samples which may lead to hallucination. [63] proposes to detect factual hallucinations by utilizing the entity's prior and posterior probabilities according to the pretrained and fine-tuned masked language models and use it as a reward signal in RL to improve kg-referenced consistency. [64] proposes a candidate summary re-ranking technique for contrastive summarization training to improve both faithfulness and summary quality. [65] use Information Extraction (IE) in a multi-task training manner to improve factual consistency of multi-document summarization. CLIFF [66] applies contrastive learning to abstractive summarization by designing four types of negative sample generation strategies to resemble errors made commonly by state-of-the-art models. CLIFF is the most related to our EnergySum because both are training discriminators on top of decoders with NCE loss. The differences lie in the structure of discriminators, the training loss, and the inference process.

Consistency improvements are often achieved at the cost of significantly lowering ROUGE scores. Therefore, correction-based methods are proposed for mitigating the tradeoff. [67] proposes a post-editing corrector module, which is pre-trained on artificial examples that are created by applying a series of heuristic transformations on reference summaries. Span-Fact is a factual correction model that leverages knowledge learned from Question Answering models to make corrections in system-generated summaries via span selection [68]. [69] proposes a fact-aware summarization model (FASum) to extract and integrate factual relations into the summary generation process via graph attention and a factual corrector model (FC) to automatically correct factual errors from summaries generated by existing systems.

Improving consistency for text summarization can benefit diverse downstream applications such as human interaction understanding [70, 71], misinformation detection reasoning [72, 73], situation report generation [39], content recommendation [74], and code understanding [75].

2.2.3 Handling Long Inputs

We introduce four main categories of methods to improve the scalability of long-input summarization: sparse attention methods, hierarchical/graph methods, divide-and-conquer methods, and retrieve-then-summarize methods.

Sparse Attention. Sparse attention methods reduce the computational overhead and enable longer inputs for text summarization models with sparse or blocked attention patterns. Sparse attention usually requires less cumbersome model structure modification and is compatible with pretrained models as a plug-in model component. However, due to the limitation of the receptive field in the sparsity mechanism, sparse attention might lose low-level

details and the dependency among semantic units and also waste the benefits of pretraining.

Sparse attention can be classified into content-independent and content-dependent. Content-independent attention such as Local Attention, Window Attention, Global Attention, and Dilated Attention work under the assumption that spatially proximate context is more important. [5] proposes a head-wise positional stride attention for encoder-decoder attention. Content-dependent attentions, such as [76], dynamically select salient tokens or sentences to constrain the attention computation. Content-independent and content-dependent are combined in [77], where local self-attention and explicit content selection collaboratively address the long-span dependencies.

Hierarchical/Graph Methods. Hierarchical encoding and graph-based methods assume particular underlying document structures and capture the semantics and dependencies in multiple levels at the cost of extra memory and time usage.

Graph structure efficiently captures inter-sentence semantics by learning multi-granularity nodes and edge representations. Hetformer[78] constructs a heterogeneous graph for long extractive summarization using multi-granularity sparse attentions. By typing the nodes into tokens, entities, and sentences, and typing the edges into token-to-token and token-to-sentence, the graph structure of the raw document can be preserved. The sparse attention patterns limit the interactions between different types of nodes.

Hierarchical models assume hierarchical structures of the documents to capture long-range dependency while preserving details. [79] is the first to hierarchically capture the document structure for abstractive long document summarization with word-level and sentence-level RNNs. [80] combines section-level and sentence-level information. [81] combines bottom-up and top-down inference for token representation learning. [82] introduces trainable hierarchical biases, which adjust attention weights based on tokens' relative positions with regard to the document tree structure. Combining hierarchical and graph structures, [83] assumes a two-level hierarchical graph representation of the source document, and proposes an unsupervised graph-based ranking model using asymmetrical positional cues to determine sentence importance.

Divide-and-conquer Methods. Divide-and-conquer methods truncate the documents into processable trunks and merge the individually processed units into a final summary. They have the advantages of easy implementation and reasonable performance but at the cost of breaking the dependency from truncated units. [84] groups input into pages by spatial, discourse and document locality, encodes each page individually, and combines the local predictions into final outputs. [85] generates a coarse summary in multiple stages and then produces the final fine-grained summary.

Retrieve-then-summarize Methods. Similar to extract-then-generate methods, Retrieve-

then-summarize methods query the documents to extract useful information with prior knowledge, such as token positions and keywords, then summarize the retrieved information. The retrieval process is implicit in vanilla transformer models, where attention weights coalesce to key positions. Explicit retrieval is more common in extractive summarization methods [86, 87], while [88] jointly trains the extractor and the generator and keeps the extracted text snippets latent for abstrative summarization.

2.3 DOCUMENT-LEVEL EVENT ARGUMENT EXTRACTION

One of the major steps in document understanding is accurate document-level information extraction, which includes Name Entity Recognition, Entity Linking, Event Extraction, Event Argument Extraction, etc. In this proposal, we focus on document-level event argument extraction.

The previous work focuses on sentence-level argument extraction approaches, where the event trigger and its arguments are usually located within a single sentence, and thus cannot handle the cross-sentence trigger-argument distribution and the existence of multiple events within one document. Though recent attempts at document-level argument extraction have gone beyond sentence boundaries, they either focus on the one-event-per-document setting or model each event independently.

The methods in [89] and [90] are designed for Role-filler Entity Extraction (REE) task under the assumption that one generic template is produced for each document, while our work focuses on extracting arguments for multiple events for each document. [91] introduces Parallel Prediction Network that generates all possible events in parallel based on the document-aware representations, while we adopt a generative framework. [92] models the whole document as graphs and captures the interdependency among events by tracking the extracted events with a global memory, while we introduce event awareness for interdependency without external memory modules. [93] introduce syntactic shortcut arcs to enhance information flow in a graph-based model for jointly extracting multiple event triggers and arguments. [16] introduces RAMS dataset for multi-sentence argument mention linking but only annotates one event per document. [94] proposes to learn within-event sentence structures for jointly extracting events and entities within a document context. The key difference between these efforts to ours is that we focus on a more challenging and more practically useful setting: consistently extracting arguments of multiple events within one document. The most related work to ours is [95], which formulates the task as conditional generation following event templates and extracts arguments for each event independently, while our work focuses on the consistency among arguments for different events.

2.4 LONG SEQUENCE PROCESSING

2.4.1 Attention Approximation

One of the bottlenecks of applying transformers [96] to long document encoding lies in the self-attention mechanism, which is known to be resource-intensive with quadratic time and space complexity ($O(n)$ where n is the input sequence length). Consequently, Transformers cannot support long sequence processing and large batch size with limited resources.

Among all the transformer acceleration methods, including attention layer simplification by pruning redundant attention heads [97, 98] and model size reduction with knowledge distillation [99, 100, 101], we focus on attention approximation models, which can be roughly classified into pattern-based and low-rank methods.

To reduce the time and space complexity by avoiding exhaustive computation over the attention metric, recent studies propose to apply sparse attention patterns to limit the number of elements participating in matrix multiplications. BlockBERT [102] introduces sparse block structures into the attention matrix. Sparse Transformer [103] introduces dilated patterns. Big Bird [104] proposes a combination of random, window, and global attention. Longformer [15] combines local windowed attention with task-motivated global attention. Informer [105] allows each key to only attend to the Top-u queries under the Kullback-Leibler divergence based sparsity measurement. Beyond limiting the attention to fixed patterns, some approaches learn the patterns by determining token assignments to relevant groups [106, 107].

Low-rank attention matrix approximation methods are based on the assumption of low-rank structure in the full self-attention matrix. Linformer [108] compresses the size of key and value matrix by Johnson–Lindenstrauss transform [109]. Performer [13] recognizes the attention score matrix as an empirical Gaussian kernel matrix and constructs a low-rank projection for both query and key matrix. Reformer [106] applies locality-sensitive hashing (LSH) [110] to simplify the computation of the attention score matrix. Synthesizer [111] aims to modify the original self-attention by replacing the dot product before softmax with Synthetic Attention, which generates the alignment matrix independent of token-token dependencies.

The work most related to Skyformer is Nyströmformer [112], which utilizes Nyström method [113, 114] instead of remodeling self-attention to approximate the attention score matrix. However, Nyströmformer applies the Nyström method to a non-PSD matrix, and thus fails to utilize the full potential of the Nyström method. This issue is resolved in our proposed method by instead lifting the kernelized attention score matrix into a large PSD

matrix which contains the target non-PSD matrix as its off-diagonal block. The work most related to Skeinformer is Informer and Linfomer, which we unify into the same sketching framework. For more details on attention approximation methods, we refer readers to a survey paper on efficient transformers [115].

2.4.2 Evaluating Efficient Transformers

Recent studies have built comprehensive evaluation benchmarks for efficient long transformers but with different focuses. Long Range Arena (LRA) [116] is the first benchmark to evaluate the long-range modeling abilities of the long transformers with tasks ranging from math operation calculation to image classification. However, some data involved is intentionally constructed to be long-input, such as the impractical byte-level sentiment classification (which is usually token-level), and considered synthetic probing for model evaluation. In addition, LRA only considers encoder self-attention, and the evaluation metrics are mostly focused on classification accuracy regardless of other aspects. Comprehensive Attention Benchmark (CAB) benchmark [26] extends the attention scope from encoder self-attention to a fine-grained attention taxonomy with four distinguishable attention patterns, namely, noncausal self, causal self, noncausal cross, and causal cross attentions. CAB collects seven real-world tasks from diverse fields of computer vision, natural language processing, speech processing, and time series forecasting. SQuALITY [117] is a challenging benchmark for long-context text generation models with question-focused summaries built on short stories. SCROLLS [118] is a suite of long-input tasks covering summarization, question answering, and natural language inference.

CHAPTER 3: CONSISTENT META-REVIEW GENERATION

In scientific opinion summarization, the synthesized opinions in the generation meta-review should be consistent with the comments from individual reviews and the final decision.

3.1 MOTIVATION

Scientific Opinion Summarization provides a succinct synopsis for scientific documents and helps readers recap salient information and understand the professional discussion. Current work on Opinion Summarization is mostly for product reviews [41, 42, 43, 48] and aims at identifying representative and consensus opinions on each aspect of interest under the assumption that the input opinions are non-controversial. However, summarizing scientific opinions is more controversial and complicated. Scientists voice agreement or disagreement for specific reasons, whereas majority voting does not always accompany consensus. Scientific meta-review summarizes the *controversies* and *consensuses* in the reviews and makes decisions.

Furthermore, most opinion summarization datasets in the product review domain for abstractive summarization systems are synthetic, redundant cut-and-paste extracts built by combining extracted snippets, or sampling a review from the collection and pretending it to be a gold-standard meta-review [48]. Meanwhile, opinion summarization in scientific domains remains less explored.

To address this gap, we introduce a new task of **Scientific Opinion Summarization**, where the output meta-reviews discuss the opinions in the input reviews and accordingly make decisions. Taking research paper meta-review generation as a typical scenario, we build the **ORSUM** dataset by collecting open-sourced paper reviews and meta-reviews from OpenReview (`https://openreview.net/`), covering 10,989 meta-reviews and 40,903 reviews from 39 conference venues. Compared to the synthetic datasets from product review domains, ORSUM is built upon large-scale real-world data, enabling the applications of supervised abstractive summarization methods and more fine-grained textual analysis. In addition to meta-review generation, the structured content of ORSUM, including ratings on different aspects and multi-turn discussions, will benefit a wide range of related tasks, such as review generation [33], recommendation prediction [119, 120], review rating prediction [121, 122], and argument pair extraction [123], and peer review assignment, reviewer confidence assessment

The task of Scientific Opinion Summarization presents a distinct set of challenges, includ-

Domain	Reviews	Meta-reviews
Product	I love these protein bars in the vanilla flavor. They taste like Rice Krispies treats with vanilla frosting ... || Nugo bars are great for breakfast, lunch or a snack ... Eat them with a tall glass of water and they will keep you satisfied for hours. || ...	These bars are fantastic and taste great like a Rice Krispy treat. Good for morning, lunch or afternoon snack and a good way to get your protein in-take. They keep you full for a long time especially if you are out and about ...
Paper	It is unclear why this work is needed. Why not use ... || The paper is well written and the math seems to be sound ... The empirical evaluation of the method is not overwhelming ... || The work appears to be sound ...	Two of the reviews suggest that the technical aspects of the paper are sound, while one reviewer questions the need for the proposed approach ... While some reviewers raised concerns about ... the majority of reviewers acknowledge the ... In light of these findings, I recommend rejection ...

Figure 3.1: Product meta-reviews and paper meta-review have different compositions: A product meta-review presents the most prominent opinion instead of summarizing opinions, while a paper meta-review summarizes different opinions and makes recommendations.

ing decision consistency, comprehensive discussion involvement, and extensive evaluation requirements. (1) *Consistency in decision guidance*: Meta-review aligns with a decision, which guides the opinion selection and discussion in the meta-review. The generated scientific meta-reviews should be able to reflect the decisions. (2) *Comprehensiveness in opinion discussion*: Unlike product meta-reviews that rely on majority voting, scientific meta-reviews access both the pros and cons, as well as opinion agreement and disagreement, to evaluate the paper from the perspective of a more senior reviewer. (3) *Extensiveness in evaluation*: The assessment of a successful meta-review should explore discussion involvement, opinion soundness, and decision consistency.

To tackle the first and second challenges, we propose a Checklist-guided Iterative Introspection (CGI^2) method. CGI^2 first breaks the task into multiple steps while constantly requesting evidence to mitigate LLM's inability to follow complicated text generation instructions and their tendency to produce hallucinations. To further enhance discussion engagement, CGI^2 iteratively revises the generated meta-review based on its own feedback derived from questions in a predefined checklist. For the third challenge, we first identify the key aspects to evaluate generated meta-reviews and propose supplementary measures for this task that can be assessed using reference-free LLM-based metrics.

Our contributions include the following:

- We introduce the task of scientific opinion summarization and construct the ORSUM dataset, which contains 10,989 meta-reviews and 40,903 reviews from 39 conferences on OpenReview. It is currently the largest paper meta-review dataset.
- We propose a Checklist-guided Iterative Introspection (CGI^2) approach, which breaks down the task into several stages and iteratively refines the summary under the guidance of questions from a checklist.
- We construct a comprehensive evaluation framework for meta-review generation and assess the generation abilities of methods in different paradigms on ORSUM.

3.2 TASK FORMULATION

Given the title, abstract, and a set of reviews from distinct reviewers of one research paper, the goal of **Scientific Opinion Summarization** is to generate a meta-review summarizing the opinions in the independent reviews and make a recommendation decision.

As noted by the area chair guidance (`https://aclrollingreview.org/aetutorial`), meta-review summarizes reviews by aggregating opinions to support the decision. It entails summarizing the key strengths and weaknesses of a paper, and explicitly evaluating whether the strengths surpass the weaknesses or the reverse. The meta-review also aggregates the final opinions of the reviewers after comprehensive discussions and offers an overall evaluation.

3.3 DATASET

3.3.1 Dataset Collection and Preprocessing

We collect the ORSUM dataset for scientific opinion reviews with gold-standard meta-reviews from OpenReview. For each paper, we collect its URL, title, abstract, decision, meta-review from the area chair, and reviews from individual reviewers. We crawl 10,989 paper meta-reviews and 40,903 individual reviews from 39 conference venues. We only keep papers with meta-reviews longer than 20 tokens and exclude comments made by non-official reviewers. Considering the diverse format and naming of related data properties across venues, we unify the format to facilitate convenient access for future research purposes. We split the dataset into train/validation/test sets with 9,890/549/550 samples, respectively.

3.3.2 Dataset Comparison

We empirically compare ORSUM with existing opinion summarization datasets (or their subsets) with gold-standard summaries, including The Rotten Tomatoes (RT) [124], Copycat [44], OPOSUM [42], Yelp [45], DENOISESUM [47], PLANSUM [48], and SPACE [125].

The Rotten Tomatoes (RT) dataset [124] consists of movie critics and their editor-written one-sentence opinion consensus for 3,731 movies. RT dataset has relatively short reviews, some of which are very objective and general comments without focus on particular aspects. Copycat [44] and OPOSUM [42] annotate small reference evaluation sets for Amazon products with Amazon Mechanical Turk (AMT). Amazon dataset provides more aspect-specific reviews by describing concrete details useful for purchase decision-making. Another human-annotated set [45] from Yelp reviews has 200 AMT-annotated summaries. Compared to Amazon reviews, Yelp reviews contain more personalized experiences. DENOISESUM [47] creates a synthetic dataset from RT [124] and Yelp [45] by sampling a review as a candidate summary and generating noisy versions as its pseudo-review inputs, where reviews not reaching consensus will be treated as noise. PLANSUM [48] is another synthetic dataset from RT [124], Yelp [45], and Amazon [44] created by sampling pseudo-reviews from a Dirichlet distribution parametrized by a content planner. SPACE [125] creates a collection of human-written general summaries and aspect summaries for 50 hotels.

Abstractiveness. The percentage of novel n-grams in the meta-review counts the ratio of n-grams that do not appear in the source reviews. This metric serves as an intuitive measure of the abstractness of the summaries [126]. Table 3.1 shows that ORSUM has more novel 4-grams in meta-reviews, indicating a greater degree of content synthesis. Moreover, different from non-scientific domains, many scientific terminologies have long been practiced in particular research areas, where duplicated usage should not be considered repetitive.

Redundancy. In order to examine the presence of insightful information in the input reviews, we assess redundancy using the Normalized Inverse of Diversity (NID) score [127] This score is calculated as the inverse of the diversity metric with length normalization.

$$NID = 1 - \frac{(entropy(D)}{log(|D|)} \tag{3.1}$$

A higher NID signifies greater redundancy. As shown in Table 3.1, ORSUM exhibits lower redundancy, which can be attributed to the fact that many reviews address distinct aspects of the paper.

Dataset Positional Bias We investigate whether the ground-truth meta-reviews favor keywords in the beginning and end positions of the individual reviews. Following [128],

Dataset	Count(SRC)	Count(TRG)	Len(SRC)	Len(TRG)
RT [124]	246,164	3,731	20.57	21.4
Copycat [44]	480	180	42.63	54.33
OPOSUM [42]	600	60	43.51	67.77
Yelp [45]	3,200	200	65.25	61.15
DENOISESUM [47]	73282	837	24.32	26.45
PLANSUM [48]	249,844	869	42.81	97.2
SPACE [125]	5000	1050	34.27	54.38
ORSUM [125]	40,903	10,989	376.36	141.76
Dataset	**Collection**	**Novel 4-gram**	**NID**	
RT [124]	Human	97.10	0.1615	
Copycat [44]	AMT	89.62	0.2506	
OPOSUM [42]	AMT	85.92	0.1260	
Yelp [45]	AMT	93.26	0.1661	
DENOISESUM [47]	Synthetic	94.12	0.2270	
PLANSUM [48]	Synthetic	91.40	0.2395	
SPACE [125]	Human	90.38	0.1671	
ORSUM [125]	Human	99.89	0.1572	

Table 3.1: We compare ORSUM with existing opinion summarization datasets that contain gold-standard summaries. SRC refers to the source or input reviews. TRG refers to the target or output meta-reviews. A higher novel 4-gram score suggests better abstractiveness, while a lower NID score implies less redundancy.

we show the position distribution by recording the position of each non-stop word in the review that also appears in the summary. Each summary is split into 10 equal-length bins. Figure 3.2 shows that while most datasets favor the beginning summary words, ORSUM showcases high consistency throughout all the bins due to the nature of the task of meta-reviewing.

3.3.3 Composition Analysis

To investigate whether the human-authored meta-reviews in ORSUM have involved the pros and cons discussion, and opinion consensus and controversy discussion, we conduct a human annotation for meta-review composition.

We select 100 meta-reviews to conduct a human annotation for meta-review composition. We draw one meta-review from each venue and randomly select the others from the rest of the training set.

We ask three annotators to label the meta-review composition in two dimensions: whether the meta-review contains a detailed discussion of the paper's strengths and weaknesses, and

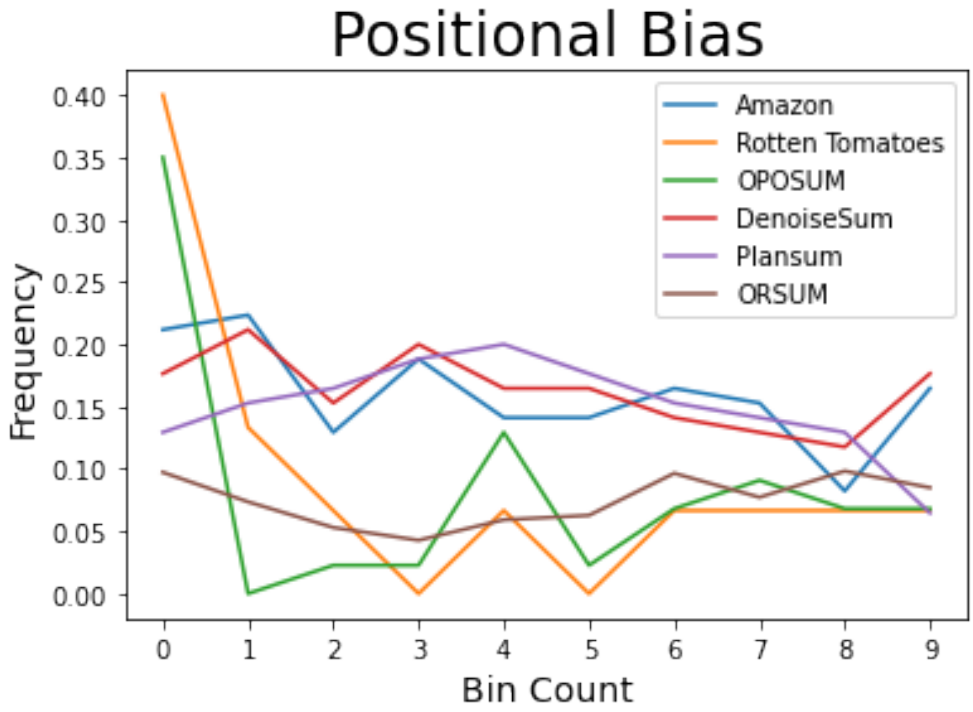

Figure 3.2: The frequency of the non-stop summary words appearing at different positions in the input. The positions are normalized to [0, 10].

whether the meta-review includes specific comments on the agreements and disagreements among the reviews. The scores range from 0 to 2, with the following interpretations: 0 indicates that the meta-review does not address the discussion at all. 1 signifies that the meta-review incorporates the discussion but lacks concrete evidence. 2 denotes that the meta-review involves a detailed discussion. For example, "The three reviewers agreed that the contribution is relevant to the workshop and presents a solid work. " is assigned a score of 1 in both dimensions because, while it refers to the discussion, the comment remains generic. The annotation process is conducted at the sentence level. If a meta-review contains a sentence with a score of 2, the entire meta-review is labeled with a score of 2.

The annotation results in Figure 3.3 reveal that only 20.7% of meta-reviews encompass both detailed discussions, regardless of their length. For each category, 47.7%, and 35.0% of meta-reviews meet the fundamental criteria for discussions of advantages and disadvantages, and consensus and controversy, respectively. Based on these results, we conclude that *the quality of human-written meta-reviews may not always be reliable.*

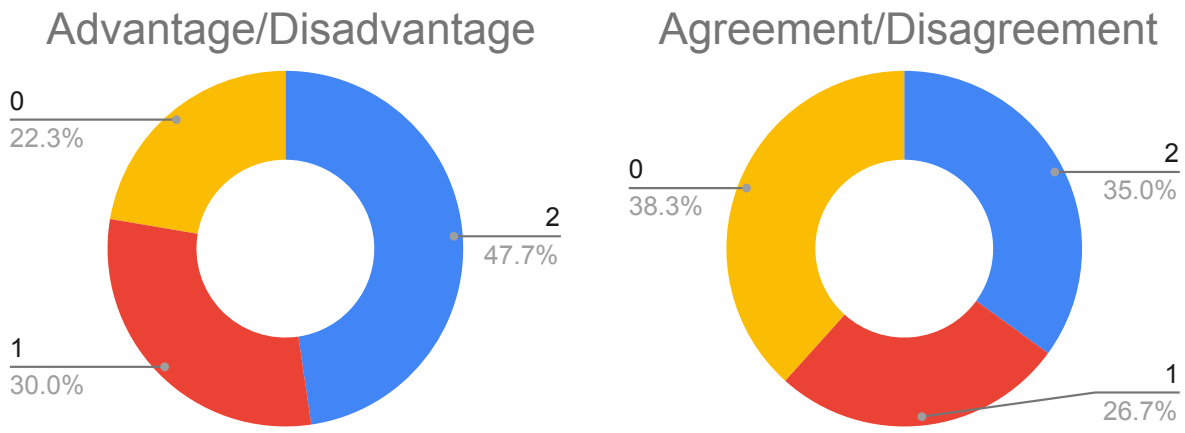

Figure 3.3: Meta-review composition. The scores range from 0 to 2: 0 indicates that the meta-review does not address the discussion at all. 1 signifies that the meta-review incorporates the discussion but lacks concrete evidence. 2 denotes that the meta-review involves a detailed discussion. Only 47.7% and 35.0% of meta-reviews meet the fundamental criteria for discussions of advantages and disadvantages, and consensus and controversy, respectively.

1. Are the most important advantages and disadvantages discussed in the above meta-review? If not, how can it be improved?
2. Are the most important consensus and controversy discussed in the above meta-review? If not, how can it be improved?
3. Is the above meta-review contradicting reviewers' comments? If so, how can it be improved?
4. Is the above meta-review supporting the acceptance/rejection decision? If not, how can it be improved?

Table 3.2: The extensible and easily adaptable checklist for Meta-review Generation accesses the essential aspects of self-consistency, faithfulness, and active engagement in discussions.

3.4 METHOD

Motivated by the unreliability of human-written meta-reviews, we turn to the training-free and reference-free prompting-based approaches. Applying Large Language Models (LLMs) like ChatGPT [129] remains a challenge due to their inability to follow complicated text generation instructions and their tendency to produce hallucinations. To address these issues, we propose to break the task into multiple steps while consistently requesting evidence. To enhance discussion engagement and evidence-based coherence in the meta-review generation, we further introduce a checklist-guided self-feedback mechanism. The process of Self-refinement [130] involves the LLM iteratively revising the generated meta-review based on its own feedback. Different from prior work, our checklist-guided self-feedback mechanism

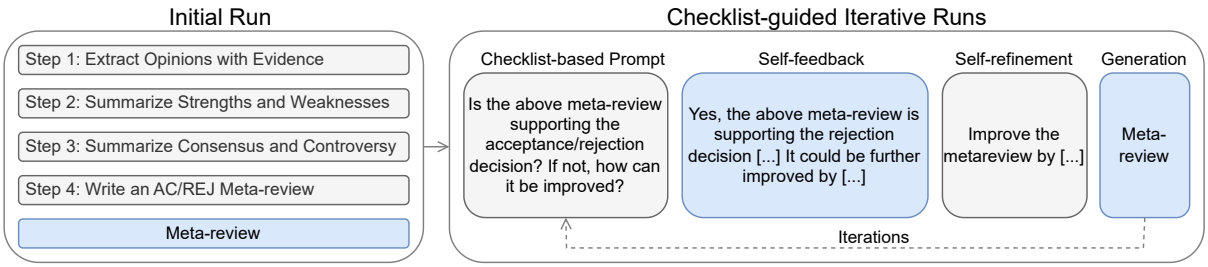

Figure 3.4: Our proposed CGI^2 framework operates through multiple iterations. In the initial iteration, the task is divided into four steps: (1) Review Opinion Extraction, (2) Strength and Weakness Synthesis, (3) Consensus and Controversy Analysis, and (4) Meta-review Drafting. For subsequent iterations, we present the black-box LLM with a query from a predefined list, acquire self-feedback, and request additional refinements.

uses self-feedback derived from questions in a predefined checklist.

Figure 3.4 illustrates our proposed Checklist-guided Iterative Introspection (CGI^2) method.

Initial Run. Given the title, abstract, and a set of reviews from distinct reviewers of one research paper, CGI^2 generates a draft of the meta-review in four steps: First, for each individual review, we prompt the LLM to extract and rank opinions and to include sentiment, aspect, and evidence. Second, based on the extracted opinions, we prompt the LLM to list the most important advantages and disadvantages of the paper and to list corresponding reviewers and evidence. Third, the LLM is prompted to list the consensuses and controversies in the above opinions and to include the corresponding reviewers and evidence. Finally, given the decision of acceptance or rejection, the LLM is requested to write a meta-review based on the above discussion.

Iterative Runs. With the meta-review draft from the initial four-step run, CGI^2 iteratively poses questions, obtains self-feedback, and requests further refinement. In each run, we first select an assessment question from a pre-constructed list of questions, as shown in Table 3.2. Customized for meta-review generation, this checklist covers the four most crucial aspects of meta-reviews. It can also be expanded and easily adapted to other complex text generation tasks. After prompting LLM with the assessment questions, we collect the refinement suggestions from the LLM's feedback. These refinement suggestions are further used as prompts for generating a revised version of the meta-review. The checklist questions are posed sequentially in one iterative run, with the number of iterations set as a hyper-parameter in CGI^2.

Our proposed approach offers two key benefits. First, it eliminates the need for external scoring functions that demand training data or human annotations. Second, it provides a general solution for employing GPT as a black box in complex text generation tasks.

3.5 EVALUATION

Meta-review generation requires a system to accurately summarize opinions, highlight reviewer consensuses and controversies, offer judgments, and make recommendations. The task complexity thus requires an evaluation that is multifaceted and goes beyond n-gram similarity. However, current evaluation metrics for long text generation are inadequate for measuring the particular requirements of meta-review generation. To address this gap, we propose a comprehensive evaluation framework that combines standard evaluation metrics with LLM-based evaluation metrics.

3.5.1 Standard Metrics

We apply standard metrics in natural language generation to assess relevance, factual consistency, and semantic coherence. For relevance, we use ROUGE-L [131] and BERTScore [50]. For factual consistency, we use FACTCC [132] and SummaC [133]. For discourse coherence, we apply DiscoScore [134]. We average the scores from these six metrics as the coherence indicator.

ROUGE-L [131] quantifies the similarity between the generated and reference texts by calculating Longest Common Subsequence. ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is one of the standard automatic metrics for summarization that calculates the overlap of a summary against a reference (usually human-produced) summary. ROUGE-N refers to the overlap of n-grams while ROUGE-L refers to the overlap of the Longest Common Subsequence (LCS). Compared to ROUGE-N, ROUGE-L takes into account the sentence-level structure similarity. In practice, we use ROUGE-L F1, which is calculated using the union LCS matches. Given a reference summary r of u sentences containing m tokens and a candidate summary c of v sentences containing n tokens,

$$R_{ROUGE-L} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, c)}{m} \tag{3.2}$$

$$P_{ROUGE-L} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, c)}{n} \tag{3.3}$$

$$F1_{ROUGE-L} = \frac{2R_{ROUGE-L}P_{ROUGE-L}}{R_{ROUGE-L} + P_{ROUGE-L}} \tag{3.4}$$

BERTScore [50] offers a more nuanced relevance evaluation as it leverages the contextualized embeddings from BERT [135] without relying on n-gram overlaps. BERTScore computes the similarity of two sentences as a sum of cosine similarities between token em-

beddings. Specifically, each token is matched to the most similar token in the other sentence to compute the Recall, Precision, and F1. For Recall, each reference token x is matched to each candidate token $\hat{x}$, while it is the versa for Precision.

$$R_{BERTScore} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \tag{3.5}$$

$$P_{BERTScore} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \tag{3.6}$$

$$F1_{BERTScore} = \frac{2R_{BERTScore}P_{BERTScore}}{R_{BERTScore} + P_{BERTScore}} \tag{3.7}$$

FACTCC [132] checks whether a given claim in the generated text is consistent with the facts presented in the source document. The BERT-based detection model is trained with an artificial dataset for three tasks: sentence consistency detection, supportive span extraction, and inconsistency span extraction.

SummaC [133] utilizes sentence-level natural language inference models for inconsistency detection. We use the convolution-based variant instead of the zero-shot variant because of the latter's sensitivity to extrema and lower correlation to human judgment. SummaC$_C ONV$ uses a 1-D convolutional layer trained on a binned NLI Pair Matrix for a (document, summary) pair.

DiscoScore [134] presents several BERT-based model variants to measure discourse coherence. The DS-FOCUS (Focus Difference) variant utilizes the theory that focus transition patterns are indicative of text coherence and models focus with contextual embeddings. The DS-SENT (Sentence Graph) variant is a graph-based method detecting whether two sentences are continuous by sharing focus. In the above methods, focus refers to noun (NN) or semantic entity (Entity).

3.5.2 LLM-based Metrics

The aforementioned methods do not evaluate discussion engagement or evidence-decision consistency. Some reference summaries may not include discussions or utilize evidence to substantiate decisions. To address this, we propose supplementary measures for this task that can be assessed and quantified using reference-free LLM-based metrics. We aim to assess the following key aspects:

- Discussion Involvement: whether the meta-review discusses the paper's strengths and weaknesses, as well as agreements and disagreements among reviewers.

- Opinion Faithfulness: whether the meta-review contradicts reviewers' comments.
- Decision Consistency: whether the meta-review accurately reflects the final decisions.

Since our requirements cannot be described as simply as one word, we explore GPT-based evaluators other than GPTScore [136]. G-EVAL [137] assesses the quality of NLG outputs by utilizing chain-of-thought (CoT) and a form-filling paradigm and has shown a very high correlation with human-based judgments. G-EVAL uses carefully constructed instructions for GPT models to follow, which subsequently yields a rating on a Likert scale ranging from 1 to 5. Likert scale scoring with ChatGPT (GPTLikert), a human-like automatic evaluation method introduced in [138] that also outperforms many standard metrics in human correlation, follows a similar evaluation protocol. These methods have shown better human alignment on multiple text summarization tasks. We are the first to adapt these methods to meta-review generation by modifying the prompts as shown in Table 3.8.

3.6 EXPERIMENTS

In this section, we aim to validate the effectiveness of our proposed method and investigate the performance of different genres of methods on the ORSUM dataset.

3.6.1 Baselines

We compare our proposed CGI^2 method with methods in different paradigms.

Abstractive Methods. PlanSum[48] uses a Condense-Abstract Framework, where reviews are condensed and used as input to an abstractive summarization model. OpinionDigest [43] extracts opinions from input reviews and trains a seq2seq model that generates a summary from a set of these opinions. MeanSum [45] is an unsupervised multi-document abstractive summarizer that minimizes a combination of reconstruction and vector similarity losses. LED [15] is a Longformer variant supporting long document generative sequence-to-sequence tasks.

Extractive Methods. LexRank [139] is an unsupervised extractive summarization method that selects sentences based on centrality scores calculated with graph-based sentence similarity. MemSum [140] models extractive summarization as a multi-step episodic Markov Decision Process of scoring and selecting sentences.

Prompting Methods. 3Sent [141] applies a simple prompt "Summary of document in N sentences" where N = 3. TCG [46] explores a four-step generation pipeline involving topic classification, sentence grouping by topic, generating chunk-wise summary per aspect, and

generating a summary per aspect. We also explore In Context Learning (ICL) [142], where a highly rated meta-review alongside the reviews is given as part of a prompt to the model. This metareview is manually picked based on adherence to the checklist mentioned above and is chosen for its fulfillment of all the criteria that define a high-quality metareview. Vanilla uses "Generate a metareview" as the prompt.

3.6.2 Implementation Details

Due to the input length constraint, each review is truncated to 300 tokens. For iterative runs in CGI^2, given the number of instructions, the reviews are deleted from the appended messages, and only discussion of these reviews with the respective evidence and initial metareview are passed forward. Similar truncation is done in the prompting-based evaluators.

For a fair comparison, all prompting methods are initiated with the gpt-3.5-turbo model. For LED we use the LEDforConditionalGeneration model from Huggingface. For MeanSum and OpinionDigest, we use their provided pretrained models. We train the content induction model of Plansum on ORSUM. In CGI^2, we set the number of iterations to 1. We show the used prompts in Table 3.6.

3.6.3 Automatic Evaluation

Higher standard metric results indicate better summarization, but not necessarily better opinion summarization. ROUGE-L, BERTScore, SummaC, and DiscoScore do not consider the multifaceted nature of meta-review, which goes beyond summarization. Our method performs near average in BERTScore and SummaC and the highest in ROUGE-L and DiscoScore amongst the prompting baselines. When compared to extractive and abstractive methods, our method performs lower since some of them specifically account for maximizing semantic similarity.

Evaluators like G-Eval and GPTLikert favor specific dimensions given in their prompts. Our method shows promising results in both G-Eval and GPTLikert due to the carefully constructed and revised prompts. Most prompting methods also outperform extractive and abstractive methods.

Human meta-reviews in the dataset scored amongst the lowest in all categories, signifying the unreliability of some human-written meta-reviews and the need for the automatic writing auxiliary process. When comparing for semantic similarity, extractive methods outperform both abstractive and prompting methods with the exception of Plansum. This is due to

Model	Informativeness	Soundness	Self-consistency	Faithfulness
Human	0.71	0.68	0.67	-
LED-finetuned	0.56	0.46	0.21	0.73
LexRank	0.87	0.94	0.16	-
CGI^2 (ours)	**0.98**	**0.92**	**0.84**	**0.79**
w/o Iterative Runs	0.97	0.76	0.48	0.74

Table 3.3: Human annotation results on meta-reviews for 50 challenging papers from the test set.

the nature of content planning in Plansum which is very central to the task of meta-review generation.

3.6.4 Human Evaluation

We conduct a human annotation on 50 challenging boundary papers from the test set, which have average review scores on the borderline of acceptance. Five anonymous baseline outputs from Human, LED-finetuned, LexRank, CGI^2, and CGI^2 without iterative runs, are shown to three annotators. The annotators are asked to provide binary labels of informativeness, soundness, self-consistency, and faithfulness for each meta-review. Informativeness measures whether the meta-review involves both strength and weakness discussion. Soundness examines whether the meta-review provides evidence to support the discussed strength or weakness. Self-consistency indicates whether the recommendation decision is clearly written and consistent with the comments in the meta-review. Faithfulness evaluates whether the meta-review contains hallucinations. We assume Human and the extractive LexRank have perfectly faithful summaries.

Results shown in Table 3.3 validate the effectiveness of our proposed method. The extractive method (LexRank) is easily biased toward one reviewer, involving no discussion nor decision, but having no hallucination problems. The abstractive method (LED-finetuned) learns to copy the sentences in the input and form a short meta-review with little discussion and sometimes internal hallucinations or repetitiveness. Our prompting-based method presents less hallucination with the evidence requirements in designed prompts. Compared to human-written meta-reviews, all automatic methods are less capable of generating in-depth analysis, which calls for knowledge enhancement.

We also observe that hallucinations in LLM are more likely to happen in summarizing consensus and controversy, which requires information integration. In contrast, hallucinations in the extractive-alike abstractive method are more likely to be triggered by generating some

general comments. Hallucination detection in scientific opinion summarization remains an opening problem.

3.6.5 Case Study

Figure 3.5 presents the meta-reviews from human, vanilla, CGI^2, and CGI^2 without iterative runs for a random paper (`https://openreview.net/forum?id=9GXoMs__ckJ`). Additionally, we show three generated examples in Table 3.7.

From the qualitative results, we have the following observations: (1) The hallucination problem is alleviated in CGI^2 because the model is constantly asked for evidence. (2) The language style of always providing a summary at the end brings redundancy in CGI^2. (3) The vanilla prompting baseline usually does not make recommendations and involve discussion, as the model fails to fully understand the complex task requirement. (4) Iterative refinement sometimes improves the concreteness of opinion discussion. However, there are two problems with the iterative refinements. First, the suggestions provided by the large language model are usually generic and less useful for further refinement. Second, more self-refinement iterations bring heavier forgetfulness for the initial instructions on opinion extraction and discussion. Table 3.4 shows the example meta-reviews generated with and without one turn of self-refinement.

3.7 CONCLUSION

In this paper, we introduce the task of scientific opinion summarization, where research paper reviews are synthesized into meta-reviews. To facilitate this task, we introduce a new ORSUM dataset, an evaluation framework, and a Checklist-Guided Iterative Introspection approach. We conduct an empirical analysis using methods in different paradigms. We conclude that human-written summaries are not always reliable, and the combination of task decomposition and iterative self-refinement shows promising discussion involvement ability and can be applied to other complex text generation using black-box LLM.

Direct extensions of this work include the incorporation of author rebuttals into the input data to enhance the model's ability to generate more balanced meta-reviews and introduce an effective and efficient hallucination detection tool for long scientific opinion summarization.

Also, the proposed meta-review generation system can be turned into an interactive tool by integrating it with a user-friendly interface in a chatbot style. The advantage of an interactive form is that it allows more user control and customization over the generated content in the middle steps. For example, opinions in the current systems are weighted

equally, while expert judgment over the selection of important aspects of the paper will be more accurate for a fair assessment. After opinion extraction and discussion synthesis steps, the experts can rank the opinions and discussions, and only keep the crucial ones that form the decision. In the meantime, the users, or area chairs, can have their own judgments and opinions, which can be inserted into the responses before the whole draft is formed.

This work on scientific opinion summarization has limitations in terms of data scope and task configuration. As the dataset is collected from OpenReview, the majority of meta-reviews are in the Machine Learning area, and many papers have been accepted. Conclusions drawn from this data distribution might not be applicable to datasets in other domains. Furthermore, to simplify the task setting, author rebuttals have not been included as input, which may also constrain the extent of discussion engagement in generating meta-reviews.

BEFORE Based on the reviews, there is a consensus that the proposed logit anchoring method is simple and effective in improving the consistency of backdoored models on clean data. The theoretical analysis also provides interesting insights into injecting backdoors with AWP. The paper also proposes a novel concept of consistency in backdoor learning and evaluates it with several metrics. However, there are some weaknesses, such as the lack of a comprehensive literature review on the consistency issue in backdoor learning and the need for more discussion on the reasons why backdoored models with AWPs are hard to detect or mitigate. Overall, the strengths of the paper outweigh the weaknesses, and the proposed method and analysis make a valuable contribution to the field of backdoor learning. Therefore, I recommend.
AFTER The paper proposes a logit anchoring method to improve consistency on clean data in backdoor learning. The authors provide theoretical analysis and empirical results that validate the effectiveness of their method. The paper also introduces a novel concept of consistency in backdoor learning and evaluates it with several metrics. The reviewers generally agree that the proposed method is simple and effective. The theoretical analysis provides interesting insights into injecting backdoors with AWP, and the experiments are solid. However, the paper lacks a comprehensive literature review on the consistency issue in backdoor learning, and there is a need for more discussion on the reasons why backdoored models with AWPs are hard to detect or mitigate. Overall, the strengths of the paper outweigh the weaknesses. Therefore, I recommend accepting the paper with minor revisions. The authors should provide a more comprehensive literature review and discuss the reasons why backdoored models with AWPs are hard to detect or mitigate. Additionally, they could simplify the presentation of the method and clarify the results in Table.
BEFORE Based on the reviews, it is clear that the paper provides important insights into the possibility and impossibility of learning from poisoned data in the presence of malicious users. The paper's theoretical contributions are significant and represent an important step forward in understanding adversarial machine learning. However, there are some weaknesses that need to be addressed. The paper contains several typos that make the reading challenging at times, and some reviewers have suggested that the results are not particularly surprising or that the paper could benefit from additional experiments or a more detailed analysis of the threat model. Overall, the strengths of the paper outweigh its weaknesses, and I believe that the paper represents an important contribution to the field of adversarial machine learning. Therefore, I would recommend accepting the paper
AFTER This paper provides important insights into the possibility and impossibility of learning from poisoned data in the presence of malicious users. The paper's theoretical contributions are significant and represent an important step forward in understanding adversarial machine learning. The paper provides matching upper and lower bounds in the realizable setting, which is a significant contribution to the field. Additionally, the paper's treatment of the agnostic setting is also important, though the gap between the upper and lower bounds is larger in this case. However, there are some weaknesses that need to be addressed. The paper contains several typos that make the reading challenging at times, and some reviewers have suggested that the results are not particularly surprising or that the paper could benefit from additional experiments or a more detailed analysis of the threat model. Overall, the strengths of the paper outweigh its weaknesses, and I believe that the paper represents an important contribution to the field of adversarial machine learning. Therefore, I recommend accepting the paper.

Table 3.4: Examples of meta-reviews before and after one-turn self-refinement.

Human

The paper unanimously receives positive rates thanks to strong motivations and interesting results. As the reviews show satisfaction on the authors' feedback, the final draft needs to respect it accordingly, for example, about the limitations of this research.

Vanilla

This paper could have benefited from a clearer and more concise presentation of the main findings. It is sometimes difficult to follow the paper's chain of thought, and the authors could have presented their results in a more accessible way. The experiments are limited to a small set of environments, and it would be interesting to see how the results generalize to other tasks. Additionally, it would be useful to have a deeper analysis of why iGPT does not perform well, as this could help inform future research on how to improve image-based pretraining.

CGI^2

After carefully considering the reviews provided, there is consensus that the paper presents an important and thought-provoking analysis of representation dynamics between pretraining and finetuning in the context of reinforcement learning. Both reviewers agree that the paper addresses an important question and presents insightful experiments. One reviewer highlights the experiments as a strong point in the paper, while the other reviewer raises a concern about the paper's tendency to present hypotheses without exploring them further.
Overall, the strengths of the paper outweigh its weaknesses, and it has the potential to make a significant contribution to the field of reinforcement learning. Therefore, I recommend accepting the paper with a minor revision to address the concern raised by the reviewer. Specifically, the paper should explore the hypotheses presented in more depth to provide a more comprehensive analysis. If the authors can address this concern, the paper can make a valuable contribution to the field of reinforcement learning.
In summary, I recommend accepting the paper with a minor revision.

Initial run

Based on the reviews, I recommend accepting the paper. The paper addresses an important and under-explored research question and presents a detailed and thought-provoking analysis of the pretraining-finetuning paradigm for Transformer models in the context of reinforcement learning. The paper also identifies relevant techniques to perform analysis of representation similarity and training dynamics. While there are some weaknesses pointed out, such as the need for further exploration of some hypotheses and occasional difficulty in following the writing and presentation, the strengths of the paper outweigh these concerns. Therefore, I recommend accepting the paper with minor revisions to address the weaknesses pointed out by the reviewers.

Figure 3.5: We show the meta-reviews from human, vanilla, CGI^2, and CGI^2 without iterative runs for the same paper. The yellow background indicates hallucinated content. The green background indicates redundant content.

Models	**ROUGE-L**	**BERTScore**	**FactCC**	**SummaC**
Human	-	-	0.538	0.368
Abstrative Methods				
PlanSum	**0.465**	0.785	0.608	0.533
OpinionDigest	0.124	0.838	0.612	0.575
MeanSum	0.132	0.827	0.559	0.464
LED	0.161	0.846	0.618	0.785
LED-finetuned	0.221	0.853	0.634	0.795
Extractive Methods				
LexRank	0.433	**0.881**	**0.729**	**0.937**
MemSum	0.337	0.827	0.683	0.825
Prompting Methods				
Vanilla	0.174	0.817	0.498	0.423
3Sent	0.109	0.783	0.562	0.503
InstructPrompt	0.208	0.823	0.543	0.449
TCG	0.189	0.847	0.544	0.466
ICL	0.192	0.847	0.578	0.470
CGI^2 (ours)	0.201	0.835	0.559	0.328
CGI^2 w/o Iterative Runs	0.118	0.830	0.536	0.332

Models	**DiscoScore**	**G-EVAL**	**GPTLikert**
Human	0.74	0.731	0.607
Abstrative Methods			
PlanSum	0.911	0.731	0.608
OpinionDigest	0.862	0.762	0.618
MeanSum	0.900	0.767	0.622
LED	0.958	0.731	0.624
LED-finetuned	0.961	0.751	0.649
Extractive Methods			
LexRank	**1.256**	0.726	0.656
MemSum	0.989	0.711	0.628
Prompting Methods			
Vanilla	0.808	0.752	0.626
3Sent	0.667	0.758	0.661
InstructPrompt	0.862	0.751	0.646
TCG	0.895	0.761	0.632
ICL	0.871	0.756	0.612
CGI^2 (ours)	0.899	**0.768**	**0.673**
CGI^2 w/o Iterative Runs	0.849	0.732	0.629

Table 3.5: ROUGE-L and BERTScore assess semantic similarity with reference text. FactCC and SummaC detect factual consistency. DiscoScore measures coherence. G-EVAL and GPTLikert are GPT-based comprehensive evaluation measures for discussion involvement, opinion faithfulness, and decision consistency.

Models	**Prompts**
Vanilla	Generate a Metareview
3sent	Generate a summary of document in 3 sentences.
InstructPrompt	Imagine you are a human metareviewer now. You will write metareviews for a conference. Please follow these steps: 1. Carefully read the reviews, and be aware of the information it contains. 2. Generate a metareview based on three dimensions: 'Discussion Involvement', 'Opinion Faithfulness' and 'Decision Consistency'. Definitions are as follows: (1) Discussion Involvement: Discuss the paper's strengths and weaknesses, as well as agreements and disagreements among reviewers, (2) Opinion Faithfulness: Do not contradict reviewers' comments, (3) Decision Consistency: Accurately reflect the final decisions.
TCG	Describe the topic of each sentence in one word. Summarize what reviewers said of the paper. Summarize the summaries of the reviews
ICL	Given a pair of reviews and a metareview as an example, Generate a metareview based on given reviews. {example}
CGI^2 (ours)	From the sentiments and aspects discussed in the reviews, what are the key strengths and weaknesses of this paper? Please cite corresponding reviewers and evidence. Identify the points of agreement and disagreement among the reviewers. Please include the corresponding reviewers and evidence. Considering the key sentiments from the reviews, the identified strengths and weaknesses, and the consensus and controversy among the reviewers, write a metareview with decision of acceptance/rejection. Are the most important advantages and disadvantages discussed in the above meta-review? Are the most important consensus and controversy discussed in the above meta-review? Is the above meta-review contradicting reviewers' comments? Is the above meta-review supporting the rejection decision? If not, how can it be improved? using this discussion, write an accurate ¡200 words metareview with decision of acceptance/rejection.
CGI^2 ablation	From the sentiments and aspects discussed in the reviews, what are the key strengths and weaknesses of this paper? Please cite corresponding reviewers and evidence. Identify the points of agreement and disagreement among the reviewers. Please include the corresponding reviewers and evidence. Considering the key sentiments from the reviews, the identified strengths and weaknesses, and the consensus and controversy among the reviewers, write a metareview with decision of acceptance/rejection.

Table 3.6: Prompts we use in all prompting-based methods.

After careful consideration of the reviews and feedback, the proposed paper presents a novel method for constructing variational approximations using SMC with partial rejection control and provides empirical validation of the approach on a variational RNN. The paper also provides clear explanations of the technical details and limitations of the approach. While the notation used in the paper can be cumbersome, this is a common issue in papers on SMC algorithms. However, there are also some weaknesses identified by the reviewers, including the complexity of the new bound and its scalability with K, the need for more clarity on the connection between the variance of the estimator for the normalizing constant obtained from particle filters and the tightness of the variational bound, and confusion over the use of Bernoulli factories and dice enterprise in the paper. Despite these weaknesses, the overall consensus among the reviewers is that the paper is a valuable contribution to the field and the proposed method has potential for further development. Therefore, the paper is recommended for acceptance with minor revisions.
The paper proposes a deep reinforcement learning model for portfolio optimization that harnesses both cross-asset dependencies and time dependencies. The proposed approach is evaluated on three sets of data from the Canadian and US stock markets, and the results show that the proposed approach outperformed baselines in terms of different metrics such as annual return and Sharpe ratio. The paper presents a clear formulation of the portfolio optimization problem and a reasonable solution to deal with permutation invariance across assets. The proposed permutation invariant architecture is also effective and has potential applications beyond portfolio management. Additionally, the theoretical characterizations presented in the paper are appreciated. However, there are some concerns about the clarity of technical details, such as the description of the architecture and the RL algorithm used to optimize the policy. There are also questions about the motivation for using the Corr layer instead of popular methods such as graph neural networks or transformers. While the contribution of dealing with permutation invariance is recognized, some reviewers note that this property is not new.
The proposed visual-inertial odometry (VIO) system is a novel and effective method that achieves competitive results, according to the four reviewers' feedback. The paper presents a clear and technically sound methodology that uses differentiable Unscented Kalman Filter (UKF) to learn the covariance matrices in an end-to-end manner. The authors' approach is refreshing as it is a learning-based VIO work that is not based on deep networks. However, there are some concerns raised by the reviewers. One reviewer pointed out the lack of empirical evaluation of the model interpretability, while the others highlighted the limited evaluation based mainly on the KITTI dataset and the lack of analysis on the sufficient distance assumption. Despite these weaknesses, the paper is well-written and presents significant novelties and contributions. Therefore, the paper is recommended for acceptance with minor revisions addressing the concerns raised by the reviewers. The authors should consider providing more empirical evaluation of the model interpretability.

Table 3.7: Examples of the meta-reviews generated by our proposed CGI^2 method.

G-EVAL

You will be given one metareview written for reviews by the committee on a paper. Your task is to rate the metareview on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria: Quality of Metareview (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby the metareview should be well-structured and well-organized. The metareview should always discuss the disadvantages and advantages of a paper and have a clear scope of the accept/reject decision. The metareview should have concrete evidence from the papers reviews and concrete comments as well.

Evaluation Steps:

1. Read the reviews carefully and identify the main topic and key points.

2. Read the metareview and compare it to the reviews. Check if the metareview covers the main topic, discusses advantages and disadvantages, if the most important advantages and disadvantages discussed in the above meta-review, if the most important advantages and disadvantages discussed in the above meta-review, if the most important consensus and controversy discussed in the above meta-review, if the above meta-review contradicting reviewers' comments, if the above meta-review supporting the acceptance/rejection decision, and if it presents them in a clear and logical order.

3. Assign a score for the quality of the meta-review on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Source Text: Reviews

Metareview: Meta-review

Evaluation Form (scores ONLY): - Quality of metareview :

Likert scale scoring with ChatGPT

Imagine you are a human annotator now. You will evaluate the quality of metareviews written for a conference by giving a mean value from 1 to 5 and no other explanation. Please follow these steps:

1. Carefully read the reviews, and be aware of the information it contains.

2. Read the proposed metareview.

3. Rate the summary on three dimensions: 'Discussion Involvement', 'Opinion Faithfulness' and 'Decision Consistency'. You should rate on a scale from 1 (worst) to 5 (best) and give me an average of these scores over all aspects from 1 to 5 calculated by the mean of all aspects.

Definitions are as follows:

(1) Discussion Involvement: Whether the meta-review discusses the paper's strengths and weaknesses, as well as agreements and disagreements among reviewers,

(2) Opinion Faithfulness: Whether the meta-review contradicts reviewers' comments,

(3) Decision Consistency: Whether the meta-review accurately reflects the final decisions.

Only generate the mean rating as a number on the likert scale, nothing else.

Table 3.8: We customize the prompts in G-EVAL [137] and GPTLikert [138] for evaluating meta-review generation to assess discussion involvement, opinion faithfulness, and decision consistency.

CHAPTER 4: CONSISTENT ENERGY-BASED DOCUMENT SUMMARIZATION

The last chapter has introduced work on improving consistency in multi-document summarization, where consistency comes across documents. In this chapter, we research the single-document scenario. In document summarization, the summaries should faithfully reflect the key ideas covered in the given documents, should be factual to common knowledge or given background knowledge, and should not be self-contradictory.

4.1 CATEGORIZATION OF CONSISTENCY IN SUMMARIZATION

The terminology of consistency in Document Understanding lacks clear categorical definitions. This terminology is closely related to the concepts of faithfulness, factuality, and hallucination. We categorize consistency by reference type:

(1) **Faithfulness / Source-referenced Consistency**: An output is considered source-referenced consistent when its content is faithful to the input document. It does not have to be factually correct. If certain spans in the generated summary cannot be supported by the input document, we consider them to be hallucinated content.

There are two types of hallucinations. **Intrinsic hallucination** refers to misplaced content. The intrinsically hallucinated span/token does come from the source document but is used in the wrong way. This phenomenon comes from the incorrect input content synthesizing process. **Extrinsic hallucination** refers to the content not found in the source document. This phenomenon mainly comes from the fine-tuning process of the target domains/datasets.

(2) **Factuality / Background-referenced consistency**: An output is considered Background-referenced consistent when its content is factually correct with reference to background knowledge. It does not have to be faithful to the input document. It usually provides additional useful information. Factual hallucination refers to the content not found in the source document but is factually correct. It stands at the intersection of Source-referenced inconsistency and KG-referenced consistency. Factual hallucinations are not necessarily harmful because it is natural to integrate the author's background knowledge into the summaries and it can provide additional world knowledge for ease of understanding.

(3) **Self-referenced / Self-contained consistency**: An output is considered self-contained or interior consistent when it has no self-contradictory content. This type of consistency is particularly important for accuracy evaluation when the output is longer than a sentence.

Source document	
Oscar-winning actress Angelina Jolie is visiting *Iraq* to boost what she sees as lagging efforts to deal with the problems of 2 million "very very vulnerable" internally displaced people in the wartorn country... *More than 4.2 million* Iraqis have fled their homes, around 2 million to neighboring states, mostly Syria and Jordan...	
Consistency type	*Example summary*
Faithfulness: The text is directly inferable from the source document.	... More than 5 million Iraqis have fled homes, 2 million to neighboring states ...
Factuality: The text contains hallucinated but true content referring to world knowledge.	American actress Angelina Jolie visits Iraq to boost efforts to help internally displaced refugees...
Self-supportiveness: The text does not contain self-contradictory errors.	... 2 million Iraqis have fled to neighboring states. Another 2 million are displaced domestically inside Syria and Jordan...

Table 4.1: Example summaries with different types of inconsistency. The errors in the sample summaries are in red.

4.2 MOTIVATION

While performing well in terms of overlap-based semantics metrics like ROUGE [131] and BERTScore [50], current abstractive summarization methods often generate inconsistent content due to the inherently noisy dataset and the discrepancy between Maximum likelihood estimation-based training objectives and consistency measurements. Inconsistency content in abstractive summarization has different interpretations, including text that is not directly inferable from the source document, is not factual with respect to world knowledge and commonsense, or is self-contradictory. We formalize the categorization of consistency into **faithfulness, factuality, and self-supportiveness**. Table 4.1 illustrates different types of consistency errors.

Previous methods improve consistency in document summarization by filtering out noisy training samples [62], applying contrastive learning [66], post-editing [67], etc., with a limited scope of consistency to faithfulness. However, addressing inconsistency solely in terms of faithfulness is inadequate. Unlike extractive methods, abstractive summarization introduces new content into the summary that is not directly copied from the source document and is not necessarily irrelevant. Hence, detecting and alleviating inconsistency calls for the introduction of a larger reference corpus alongside the source document. Factuality compares the generated content against world knowledge, while self-supportiveness verifies whether the

generated sentence is consistent with its preceding one.

In addition, consistency is measured on the entire prediction sequence while existing summarization objectives evaluate conditional distributions for individual tokens and lack global control over predictions. These motivate us to apply the Residual Energy-based Model (REBM) [143] framework to document summarization, which jointly trains a summarizer and a discriminator that learns to assign high scores to consistent summaries and low scores to inconsistent ones. The advantage of the energy-based methods [144] is that they score the entire input simultaneously and avoid local normalization traps, offering a natural solution to address this issue.

Therefore, we introduce **EnergySum** that adapts the REBM framework for improving consistency. We design the energy functions that reflect each type of consistency and are agnostic to summarization model instances. We propose joint inference where energy scorers cooperate with decoding searching strategies in the candidate re-ranking step. In summary, our contributions include:

- We formalize the categorization of consistency in document summarization into faithfulness, factuality, and self-supportiveness.
- We propose the EnergySum framework, which includes consistency-constrained energy scorers and joint inference. We are the first to introduce energy-based methods to consistent document summarization.
- We conduct extensive experiments on XSUM and CNN/DM datasets to validate the effectiveness of EnergySum.

4.3 METHOD

In the proposed EnergySum framework, we design energy scorers that correlate each type of consistency and integrate energy scores in candidate re-ranking during sampling.

4.3.1 Background: Energy-Based Models

Energy-Based Model (EBM) [145] is a general learning framework that assigns an unnormalized energy score to any given input. EBM has been applied in machine translation to solve the discrepancy between the training objective (Maximum Likelihood Estimation) and the task measure (BLEU) [146], and in improving calibration in natural language understanding [144].

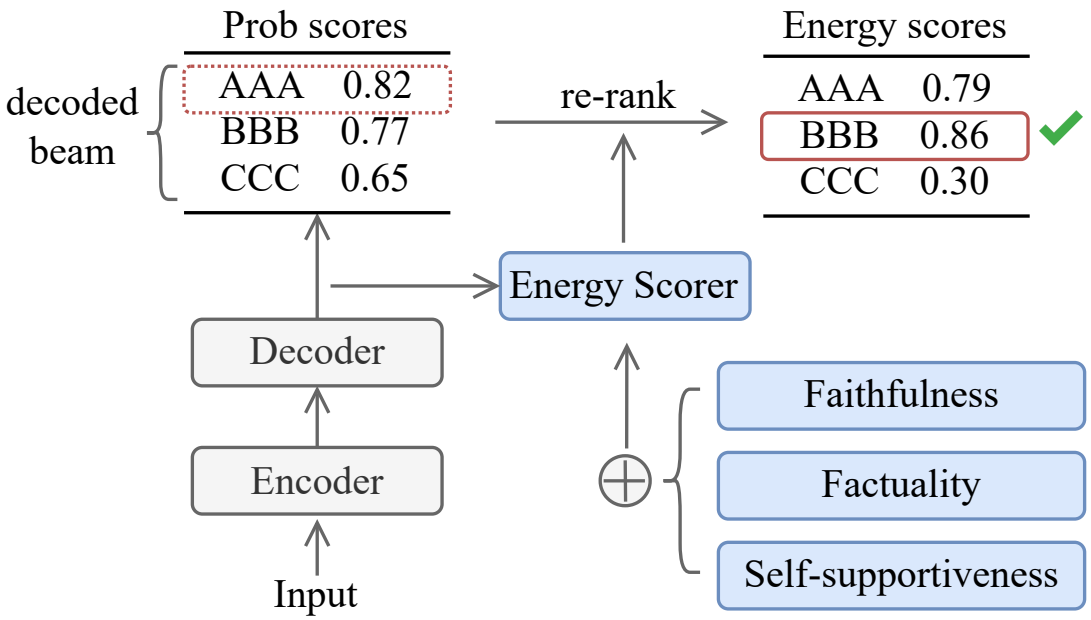

Figure 4.1: Overview of EnergySum framework. The energy scorer is a discriminator consisting of three consistency-constrained energy functions. During inference, we re-rank the decoded beam of summaries by energy scores.

Residual Energy-Based Models (R-EBMs) [143] are introduced to text generation, which use EBM to learn from the residual errors of an auto-regressive generator to reduce the gap between the model and data distributions: $P_\theta \propto P_{LM}(x)\exp(-E_\theta(x))$, where P_{LM} is a locally normalized language model and E_θ is the energy function. [147] further applies R-EMBs to end-to-end speech recognition.

Energy functions have also been used as constraints in text generation. The COLD decoding framework [148] unifies constrained generation by specifying constraints through an energy function, then performing efficient differentiable reasoning over the constraints through gradient-based sampling.

REBM vs Contrastive Learning. Though both are training discriminators on top of decoders with NCE loss, REBM differs from contrastive learning in several ways: (1) the structures of the discriminators are different, as in contrastive learning the role of the discriminator is to give a binary label while in REBM the discriminator is expected to quantify a specific phenomenon with unnormalized scores, (2) the training losses are different in implementation, and (3) the inference processes are different whereas in contrastive learning the discriminator is not involved in sampling and in our framework energy scores are used in re-ranking.

4.3.2 Energy Functions for Consistency

Energy functions solve the discrepancy between MLE-based training objectives and consistency measurements. General-purpose energy function designs are usually as simple as the mean pooling over the last encoder/decoder layer logits. To improve consistency, we propose three energy functions and use their weighted sum as the final energy function in the Noise Contrastive Estimation loss.

$$\mathcal{E}(x, y, \hat{y}) = \lambda_1 \mathcal{E}_i(y, \hat{y}) + \lambda_2 \mathcal{E}_i(x, \hat{y}) + \lambda_3 \mathcal{E}_i(\hat{y}) \tag{4.1}$$

where x is the input document, y is the reference summary, and $\hat{y}$ is the generated summary.

Faithfulness. Following [148] we use EISL (Edit-Invariant Sequence Loss) [149] as a similarity measure. This n-gram matching function can be seen as a differentiable approximation to the BLEU-n metric. Its computation is essentially a convolution operation on the candidate sequences using target n-grams as kernels.

$$\mathcal{E}_1(y, \hat{y}) = \text{EISL}(y, \hat{y}) \tag{4.2}$$

Using the reference summary to measure faithfulness benefits stable and efficient training. However, it cannot avoid dataset noise from annotation as it is based on the assumption that the reference summary is correct.

Factuality. [63] proposes to detect factual hallucinations by utilizing the entity's prior and posterior probabilities according to the pretrained and fine-tuned masked language models as classifier inputs. It is still under exploration how these two distributions work together for factual hallucinations. To apply this measure, we first initiate and freeze the pretrained BARTlarge model as the prior model. A classifier γ takes the concatenation of outputs from the prior and posterior models as its input.

$$\mathcal{E}_2(x, \hat{y}) = \gamma(p_{prior}(\hat{y}|x), p_{posterior}(\hat{y}|x)) \tag{4.3}$$

Self-supportiveness. A non-linear layer ϕ on top of the decoder outputs detects self-supportiveness in the generated summary.

$$\mathcal{E}_3(\hat{y}) = \phi(p(\hat{y})) \tag{4.4}$$

4.3.3 Training Loss

The pretrained language model is fine-tuned using the cross entropy loss $\mathcal{L}_{CE}$:

$$\mathcal{L}_{CE} = -\sum y_i \log \hat{y}_i \tag{4.5}$$

For stable and effective training of the discriminator, we combine the two squared hinge loss $\mathcal{L}_{\mathcal{E}}$ [150] and a similarity-based NCE loss $\mathcal{L}_{\text{sim}}$ [66].

$$\begin{aligned}\mathcal{L}_{\mathcal{E}} &= \mathbb{E}_{x_+}(\max(0, \hat{\mathcal{E}}_\theta(x_+)) - m_1))^2 \\ &+ \mathbb{E}_{x_-}(\max(0, m_2 - \hat{\mathcal{E}}_\theta(x_+)))^2\end{aligned} \tag{4.6}$$

m_1 and m_2 are margin hyper-parameters with which the loss function penalizes samples with energy $\hat{\mathcal{E}} \in [m_1, m2]$.

$$\mathcal{L}_{\text{sim}} = -\mathbb{E} \log \frac{\exp(\text{sim}(h_i, h_j))}{\sum \exp(\text{sim}(h_i, h_k))} \tag{4.7}$$

In the above loss, P and N are the positive sample set and the negative sample set, $y_i, y_j \in P, y_j \neq y_i, y_k \in P \cup N, y_k \neq y_i$. h_i, h_j, h_k are representations for summaries y_i, y_j, y_k, and $\text{sim}(\cdot, \cdot)$ calculates the cosine similarity between summary representation.

The final training loss is a combination of the above losses:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{\mathcal{E}}\mathcal{L}_{\mathcal{E}} + \lambda_{\text{sim}}\mathcal{L}_{\text{sim}} \tag{4.8}$$

4.3.4 Joint Inference

Previous work [143] suggests that a sample-resample procedure is similar to exact sampling from the joint distribution. Therefore, we modify the sampling process by inserting the energy scores into the candidate re-ranking step.

In decoding, a batch of sentence candidates is generated and scored for each input. We replace the generation probability scores with energy scores for the candidates and re-rank the batch. Since beam search is more likely to generate similar results, where re-ranking takes less effect, we select diverse beam search [151] as the default searching strategy. This candidate reranking strategy can be applied to other sampling methods as well. In practice, we find that diverse beam search reaches the best trade-off of accuracy and consistency.

The above-mentioned energy score involvement is done at the summary level. However, it is indeed possible to perform sentence-level or even phrase-level in the process of beam

Dataset	Train	Validation	Test	Src_len	Trg_len
XSUM	204,045	11,332	11,334	431	23
CNN/DM	287,227	13,368	11,490	781	56

Table 4.2: Dataset statistics for consistent document summarization.

search. The advantage of the more fine-grained energy score is that it might have a more accurate measurement of consistency. The disadvantage is that it is significantly more costly and it will sometimes lose long-term dependency information for some measurements.

4.4 EXPERIMENTS

4.4.1 Datasets

We conduct experiments on XSUM [152] and CNN/DM [153] datasets. XSUM (eXtreme SUMmarization) [152] comprises around 227k British Broadcasting Corporation (BBC) articles, where the first sentence of the article is treated as a summary of the article. CNN/DM (CNN/DailyMail) [153] contains over 300k news articles from CNN and the Daily Mail. The dataset statistics are shown in Table 4.2.

Data Augmentation (Positive/Negative Sample Generation). The training of our framework requires positive/negative samples to shape energy scorers. Therefore, following CLIFF [66], we adopt back-translation as the positive sample generation strategy, and syslow and entswap as the negative sample generation strategies:

- BACK-TRANSLATION: we utilize the nlpaug tool to generate positive samples.
- SYSLOWCON: We select the baseline-generated candidates in the same beam with low model confidence as the negative samples.
- SWAPENT: We construct negative samples by swapping entities in the references with other randomly selected entities of the same entity type in the source.

We show some augmented data samples in Table 4.3.

4.4.2 Baselines

. We compare our methods with the following baselines:

- Human baseline refers to the human-written reference summaries.

- BARTlarge [154] is a denoising autoencoder for pretraining sequence-to-sequence model. We finetune the pretrained BARTlarge.
- LOSSTRUNC [62]is a simple and scalable procedure that adaptively removes high log loss examples as a way to optimize for distinguishability.
- CLIFF [66] is a contrastive learning framework, which leverages both reference summaries, as positive training data, and automatically generated erroneous summaries.
- FASUM [69] is a fact-aware summarization model to extract and integrate factual relations into the summary generation process via graph attention with a factual corrector model FC to automatically correct factual errors from summaries generated by existing systems.

4.4.3 Evaluation Metrics

We evaluate accuracy with ROUGE [49] and BERTScore [50]. For faithfulness and factuality, we measure with FEQA [52] and ENTFA [63], respectively. Since there is no existing metric for self-supportiveness, we propose DAESS, which splits the multi-sentence summary and adapts DAE [57] to compare every pair of sentences in one summary. The summaries in the XSUM dataset are usually one sentence, so we only evaluate DAESS on the CNN/DM dataset.

4.4.4 Implementation Details

We instantiate EnergySum and Losstrunc both with the pretrained BARTlarge [154] model. The margin hyperparameters $m_1 = -10, m_2 = -5$ in $\mathcal{L}_{\mathcal{E}}$ are selected by performance on the development set.

For FASUM, we evaluate the provided prediction files as the code is not publicly available. Note that their provided test set file is slightly different than the standard test set split. For all other experiments, each model is trained for 15000 steps, the learning rate is set to $1e-3$, the max token in one batch is set to 4096, the update frequency is 16, and the optimizer is Adam with 500 warm up steps. The hyperparameter c in Losstrunc is set to 0.3.

For numerical consistency, all experiment results are averaged across three random runs. On average it takes approximately ten hours to train a model with one Tesla A100 GPU with 40GB DRAM. Since evaluating FEQA over the whole test set is time costly, we randomly sample 500 document-summary pairs to calculate the metrics.

4.4.5 Results and Discussion

Table 4.4 shows that EnergySum improves faithfulness with comparable accuracy performance on both XSUM and CNN/DM compared to BARTlarge. All consistency improvement baselines have lower overlapped-based accuracy than BARTlarge, showing the trade-off between MLE-based training and consistency training. Nevertheless, our method hurts less from such a trade-off and still has comparable accuracy performance.

Human-written gold summaries usually represent the upper bound of the performance. However, the human baseline has lower FEQA (faithfulness) performance, indicating the existence of noise in the dataset. Self-supportiveness scores are all close to 100%, implying that self-supportiveness is not a challenging problem for current summarization systems and also calling for a more fine-grained evaluation metric.

There is also a trade-off between the sampling method selection and the overall performance. Joint inference can only be applied to searching strategies where the searched candidates are diverse, which in general performs worse than regular beam search.

4.5 CONCLUSION

We propose to apply the Residual EBM framework with energy scorers and joint inference to improve consistency in document summarization. Experiments on XSUM and CNN/DM datasets show that EnergySum mitigates the trade-off between accuracy and consistency. Direct extensions of this work include proposing more fine-grained data augmentation strategies and investigating the relation between prediction certainty and energy scores.

This work on consistent document summarization has limitations in terms of data scope and task configuration. First, EnergySum learns from common errors simulated by data augmentation strategies, which could limit its application in more diverse contexts. Second, EnergySum predicts sentence-level scores and thus cannot detect span-level errors or predict error types.

Original	Irish Taoiseach (PM) Leo Varadkar has engaged in some "sock diplomacy" in his first meeting with Canadian Prime Minister Justin Trudeau in Dublin.
Back-translation	Irish Prime Minister Leo Varadkar has engaged in a kind of "sock diplomacy" during his first meeting with Canadian Prime Minister Justin Trudeau in Dublin.
Syslow	Canadian Prime Minister Justin Trudeau has met Taoiseach (Irish Prime Minister) Leo Varadkar in Dublin. Canadian Prime Minister Justin Trudeau has met Taoiseach (Irish prime minister) Leo Varadkar in Dublin. Canadian Prime Minister Justin Trudeau has met Taoiseach (Irish PM) Leo Varadkar in Dublin. Canadian Prime Minister Justin Trudeau has met Taoiseach (Irish Prime Minister) Leo Varadkar for the first time.
Swapent	Irish Taoiseach (PM) Leo Varadkar has engaged in some "sock diplomacy" in his first meeting with Canadian Prime Minister Justin Trudeau in Dublin. Mr Trudeau's (PM) Leo Varadkar has engaged in some "sock diplomacy" in his first meeting with Canadian Prime Minister Justin Trudeau in Dublin. Irish Taoiseach (PM) Leo Varadkar has engaged in some "sock diplomacy" in his first meeting with Irish Prime Minister Justin Trudeau in Dublin. Irish Taoiseach (PM) Leo Varadkar has engaged in some "sock diplomacy" in his first meeting with Canadian Prime Minister Sophie GrÃ© in Dublin. Irish Taoiseach (PM) Leo Varadkar has engaged in some "sock diplomacy" in his first meeting with Canadian Prime Minister Justin Trudeau in Germany.

Table 4.3: Example of positive and negative samples in the XSUM validation set.

Dataset	Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE
XSUM	Human	-	-	-	-
	BARTlarge	43.64	20.04	34.34	91.56
	FASUM	30.61	10.06	23.97	88.53
	FASUM+FC	30.53	10.00	23.89	88.58
	Losstrunc	41.73	17.88	32.68	91.24
	CLIFF	**42.07**	**18.50**	32.82	91.29
	EnergySum	41.69	18.12	**32.98**	**91.44**
CNN/DM	Human	-	-	-	-
	BARTlarge	43.86	21.07	40.74	88.70
	FASUM	40.83	17.94	37.78	88.08
	FASUM+FC	40.68	17.77	37.63	88.24
	Losstrunc	36.37	17.35	34.21	87.72
	CLIFF	42.15	19.82	38.91	87.95
	EnergySum	**43.38**	**20.45**	**40.27**	**88.27**

Dataset	Model	FEQA	ENTFA	DAESS
XSUM	Human	18.95	72.27	-
	BARTlarge	29.13	68.38	-
	FASUM	18.38	55.83	-
	FASUM+FC	19.77	54.91	-
	Losstrunc	28.94	66.31	-
	CLIFF	25.28	**83.87**	-
	EnergySum	**30.26**	68.45	-
CNN/DM	Human	30.94	91.46	99.95
	BARTlarge	18.06	63.50	99.92
	FASUM	18.75	61.23	**99.89**
	FASUM+FC	18.74	60.53	**99.89**
	Losstrunc	11.58	65.90	99.65
	CLIFF	21.33	64.90	99.86
	EnergySum	**41.92**	**66.43**	**99.89**

Table 4.4: Results(%) on XSUM and CNN/DM datasets. ROUGE and BERTSCORE indicate accuracy. FEQA, ENTFA, and DAESS evaluate faithfulness, factuality, and self-supportiveness, respectively. For all scores, the higher the better.

CHAPTER 5: CONSISTENT EVENT-AWARE ARGUMENT EXTRACTION

The above two chapters introduce consistency improvement in scientific opinion summarization and document summarization, where the outputs are unstructured text. In this chapter, we research consistency in tasks with structured output. In document-level event argument extraction, the process units are usually scattered across the long document, and the automatic information extractors are expected to represent and identify the interrelated event argument roles within a long document.

In addition to the output format, the differences between this chapter and the last two chapters are also in the types of consistency. In scientific opinion summarization, we focus on the consistency between meta-review and individual review, and the one between meta-review and its own decision. In document summarization, we generalize the underlying relationship to faithfulness, factuality, and self-supportiveness, while in event argument extraction, we put more focus on the relationship between the extracted units.

5.1 MOTIVATION

Document-level Event Argument Extraction aims at identifying arguments and their roles for multiple events in a document. It is a practically more useful but more challenging task than sentence-level Argument Extraction [155, 156, 157] because in a typical long input document events usually scatter across multiple sentences and are inherently connected.

Multiple events in one document are usually interconnected, and thus the arguments are under certain consistency constraints. In Figure 5.1, the roles of the shared argument *Ahmad Khan Rahami* in multiple events are constrained because the *Attacker* in the *DetonateExplode* event is likely to be the *IdentifiedRole* in *IdentifyCategorize* event, the *Detainee* in the *ArrestJailDetain* event, as well as the *Defendant* in *ChargeIndict* event. Motivated by the one-sense-per-discourse theory [158] that mentions of an ambiguous word usually tend to share the same sense in a given discourse, we hypothesize that **a participant tends to play consistent roles across multiple events in the same document**. However, previous work such as [89, 91, 95] on document-level event argument extraction focuses on modeling each event independently and ignores the relation between events, and thus the extracted arguments of multiple events may violate the constraints from event-event relations. We call this inconsistency phenomenon.

Though received much attention in various areas like Abstractive Summarization [58, 69, 159] and Question Answering [160, 161], the inconsistency phenomenon addressed in

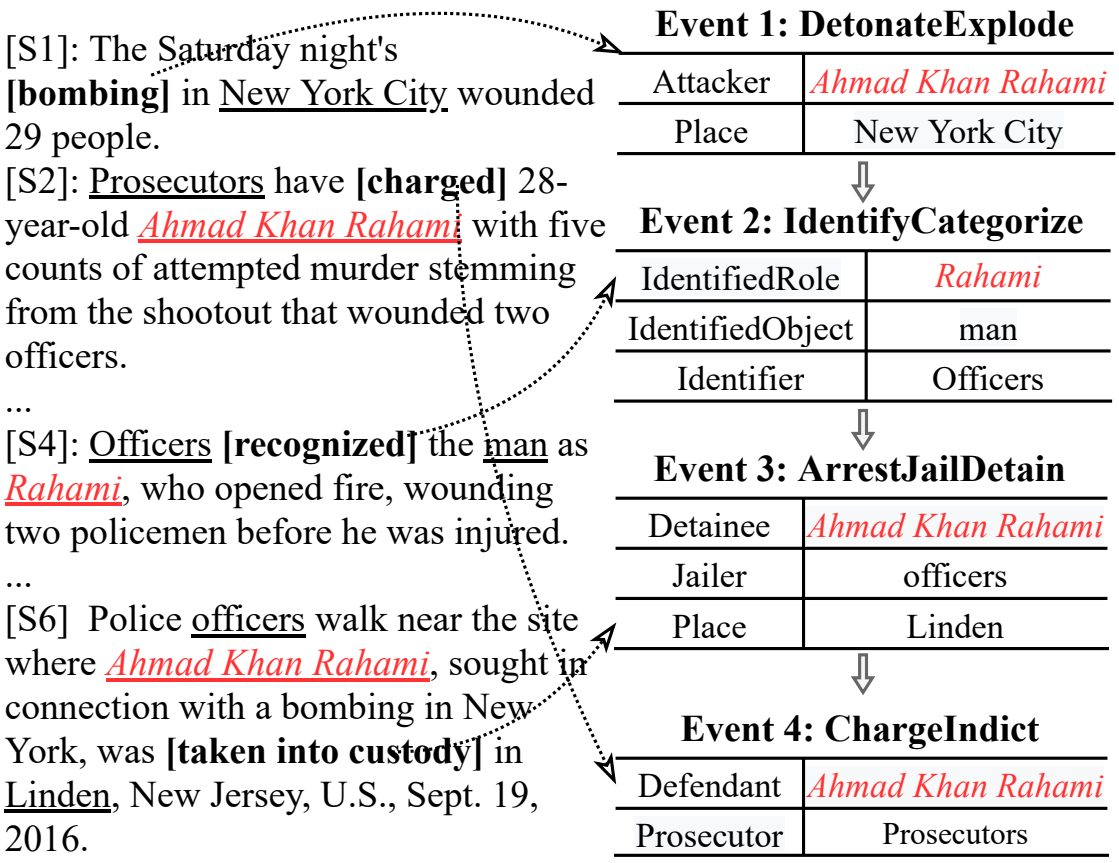

Figure 5.1: Examples of extracting arguments for multiple events in one document. The casual relation between the *Arrest* event and the *Detonate* event puts their arguments under consistency constraints: *Ahmad Khan Rahami*, the detainee in Event 3, is very likely to be the attacker in Event 1. Sentence-level models tend to miss the cross-sentence attacker argument in Event 1.

previous research focuses on factual consistency instead of self-contained consistency as in document-level argument extraction. We approach this problem with inspiration from human behavior: while reading, humans subconsciously infer the event-event relations and correctly identify the event arguments under the perceived constraints. Therefore, we refer consistent argument extraction to applying the underlying Event-Event Relations as constraints in multi-event argument extraction.

An intuitive solution to improve consistency is to incorporate explicit Event-Event Relations into the extraction process as additional input. However, the underlying event-event relations are hard to identify and classify due to the lack of reliable resources as supervision signals, especially when the arguments are unknown. In addition, precise event-event relations may not be necessary for argument extraction when the implicit connections can

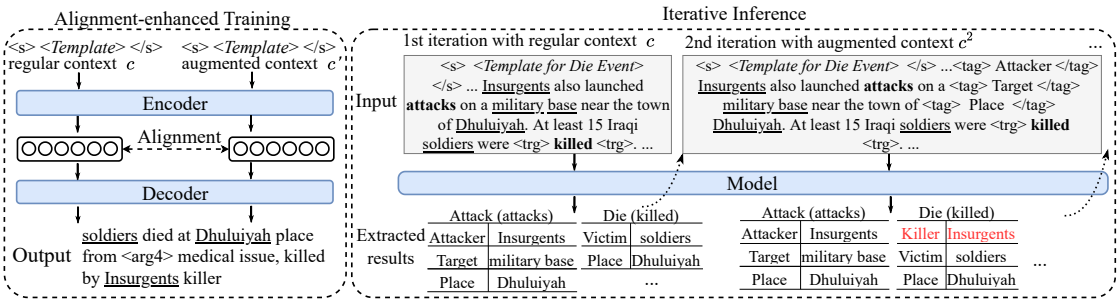

Figure 5.2: Our proposed Event-aware Argument Extraction model with alignment-enhanced training and iterative inference. During training, an auxiliary training loss aligns the event argument representations under regular context and augmented context. During inference, the context is augmented with results from the last iteration.

already well support argument extraction.

To avoid explicit modeling of event-event relations, we label the arguments of other events in the context as an implication of event-event relations. We propose an **Event-Aware Argument Extraction (EA^{2}E)** model, which incorporates alignment-enhanced training and iterative inference. When extracting arguments, the context can be self-augmented by tagging the argument labels of other events. Alignment-enhanced training implicitly introduces event awareness by pulling close the argument representation distributions under regular context and augmented context, where ground-truth argument labels of neighboring events are labeled. Iterative inference explicitly encourages event awareness by augmenting the context with the extracted arguments from the last inference iteration. The advantage of this method is that no predefined Event-Event Relation is required, nor event schema.

5.2 METHOD

Motivated by the observation that introducing event-event relations benefits the consistency of event argument extraction, we propose to incorporate implicit event-event relations with an Event-Aware Argument Extraction (EA^{2}E) model. As shown in Figure 5.2, EA^{2}E contains alignment-enhanced training and iterative inference with self-augmented context. When extracting the arguments for a target event, the context is augmented by labeling the arguments from neighboring events. During training, an auxiliary training loss pulls close the event argument representations under the regular context and self-augmented context. During inference, iterative inference encourages event awareness by using the extraction arguments from the last inference iteration as inputs.

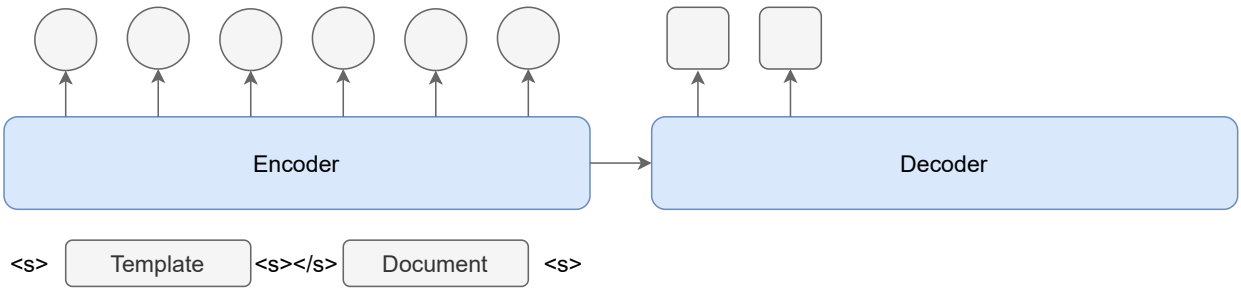

Figure 5.3: We base our model on BART, an encoder-decoder pretrained model.

5.2.1 Base Encoder-Decoder Model

Following [95], we formulate event argument extraction as a conditional generation task under the assumption that there exists a pre-defined event ontology describing each event type with an unfilled template with argument placeholders. For example, the template for *Attack* events is *<arg> detonated or exploded <arg> explosive device using <arg> to attack <arg> target at <arg> place*. Formally, given a document context c and an event trigger x with template t, the task is to extract a set of arguments $y = \{a_1, a_2, ...a_n\}$, where each a_i corresponds to a role predefined in the ontology.

We base our model on BART [154], an encoder-decoder pretrained model. Figure 5.3 shows the model structure. The input sequence is the concatenation of the document context and an event template, constructed as *<s> template </s> </s> context </s>*. The output is a filled-in template, where the tokens are all selected from the input context or template.

The model parameter θ is trained by minimizing the argument extraction loss, the conditional probability over all instances:

$$\mathcal{L}_E = -\sum log p_\theta(y|x, t, c) \tag{5.1}$$

5.2.2 Self-Augmented Context

We refer event awareness as the implication of event-event relations and reach this goal by labeling the arguments of other neighboring events in the context. Given the arguments of the neighboring events $\{j \in \mathcal{N}_i\}$, which have small token-wise distances to the target event i, we augment the regular context c by labeling them with ¡*tag*¿:

$$c'_i = \mu(c_i, \{y_j\}), \tag{5.2}$$

where μ is the tagging operation, and y_j is the arguments of event j.

For example in Figure 5.1, when extracting arguments for the target event *bombing* (Event 1), the augmented context is "The Saturday night's <trg> bombing <trg> in New York City, wounded 29 people. *<tag> Prosecutor <tag>* Prosecutors have charged 28-year-old *<tag> Defendant <tag>* Ahmad Khan Rahami...", where the two tags highlight the arguments of Event 4.

5.2.3 Alignment-Enhanced Training

An encoder is considered consistent when it is able to understand and encode the underlying relation between events into the text representations. Therefore, we propose to enhance the encoder with an auxiliary training loss $\mathcal{L}_T$ that pulls close the argument representation distributions under regular context c and under augmented context c'. During training, c' is constructed by tagging the ground-truth arguments of neighboring events.

$$\mathcal{L}_T = \sum \|p(a|c), p(a|c')\|_2 \tag{5.3}$$

The final training loss is a weighted sum of argument extraction losses ($\mathcal{L}_E$ for regular context c and $\mathcal{L}_{E'}$ for augmented context c') and alignment-enhanced loss ($\mathcal{L}_T$) with weights α and β :

$$\mathcal{L} = \mathcal{L}_E + \alpha\mathcal{L}_{E'} + \beta\mathcal{L}_T \tag{5.4}$$

5.2.4 Iterative Inference

Iterative inference explicitly introduces event awareness by utilizing extracted results in multiple inference iterations. In the first iteration, for each target event trigger i, the model obtains the predicted results y_i^1 given the regular context c_i^1. For the k-th iteration of inference, for each event trigger i the context c_i^k is augmented by labeling the extracted arguments $\{y_j^{k-1}\}$ of neighboring events $\{j \in \mathcal{N}_i\}$ from the $(k-1)$-th iteration.

$$c_i^k = \mu(c, \{y_j^{k-1}\}) \tag{5.5}$$

Model	Argument Identification					
	Head Match			Coref Match		
	Precision	Recall	F1	Precison	Recall	F1
BERT-CRF [162]	72.66	53.82	61.84	74.58	55.24	63.47
ONEIE [157]	68.16	56.66	61.88	70.09	58.26	63.63
BART-Gen [95]	70.43	71.94	71.18	71.83	73.36	72.58
EA^{2}E (ours)	76.51	**72.82**	**74.62**	77.69	**73.95**	**75.77**
w/o Alignment-enhanced Training	**77.26**	71.23	74.12	**78.61**	72.47	75.42
w/o Iterative Inference	75.96	72.29	74.07	77.13	73.42	75.22
Model	**Argument Classification**					
	Head Match			**Coref Match**		
	Precision	Recall	F1	Precison	Recall	F1
BERT-CRF [162]	61.87	45.83	52.65	63.79	47.25	54.29
ONEIE [157]	63.46	52.75	57.61	65.17	54.17	59.17
BART-Gen [95]	65.39	66.79	66.08	66.78	**68.21**	67.49
EA^{2}E (ours)	70.35	**66.96**	**68.61**	71.47	68.03	**69.70**
w/o Alignment-enhanced Training	**71.10**	65.54	68.21	**72.25**	66.61	69.32
w/o Iterative Inference	69.61	66.25	67.89	70.72	67.32	68.97

Table 5.1: Performance (%) on WIKIEVENTS dataset.

5.3 EVALUATION

5.3.1 Settings

We evaluate our proposed method on WIKIEVENTS [95] dataset and ACE 2005 English dataset (`https://www.ldc.upenn.edu/collaborations/past-projects/ace`). WikiEvents is a document-level event extraction benchmark dataset that includes complete event and coreference annotation. ACE 2005 (English subset) consists of data of various types annotated for entities, relations, and events. The dataset statistics are shown in Table 5.3.

Following previous work [163], we consider an argument span to be correctly identified when its offsets match any of the reference arguments of the current event (i.e., **Argument Identification**), and to be correctly classified when its role matches (i.e., **Argument Classification**).

Following [95, 164], we report the argument extraction performance in terms of Head Word F1 and Coreferential Mention F1. For Head Word F1, full credit will be given when the extracted argument has the same headword with the gold-standard argument. For the latter, full credit will be given when the extracted argument is coreferential with the gold-standard argument.

Model	Argument Identification					
	Head Match			Coref Match		
	Precision	Recall	F1	Precision	Recall	F1
BERT-CRF [162]	65.77	51.04	57.48	67.11	52.08	58.65
ONEIE [157]	63.33	61.46	62.38	65.12	63.19	64.14
BART-Gen [95]	70.00	73.84	71.87	71.37	75.29	73.27
EA^{2}E (ours)	**74.54**	74.88	**74.71**	**75.81**	76.16	**75.98**
w/o Alignment-enhanced Training	73.95	74.25	74.10	75.28	75.58	75.43
w/o Iterative Inference	74.36	**75.00**	74.68	75.56	**76.22**	75.88
Model	**Argument Classification**					
	Head Match			**Coref Match**		
	Precision	Recall	F1	Precision	Recall	F1
BERT-CRF [162]	56.82	44.10	49.66	57.72	44.79	50.44
ONEIE [157]	58.50	56.77	57.62	60.11	58.33	59.21
BART-Gen [95]	65.72	69.33	67.47	66.76	70.43	68.54
EA^{2}E (ours)	**71.83**	**72.16**	**72.00**	**72.98**	**73.32**	**73.15**
w/o Alignment-enhanced Training	70.78	71.06	70.92	72.05	72.34	72.19
w/o Iterative Inference	71.49	72.11	71.80	72.58	73.20	72.89

Table 5.2: Performance (%) on ACE2005 dataset.

We compare EA^{2}E with document-level BART-Gen [95], sentence-level ONEIE [157] and BERT-CRF [162]:

- **BART-Gen** [95] is a document-level neural event argument extraction model by formulating the task as conditional generation following event templates.
- **ONEIE** [157] is a sentence-level end-to-end graph-based method that extracts the globally optimal IE result as a graph from an input sentence by sentence encoding, identifying entity mentions and event triggers as nodes, computing label scores using local classifiers, and searching for the globally optimal graph with a beam decoder.
- **BERT-CRF** [162] is a sentence-level neural model by incorporating lexical and syntactic features such as part-of-speech tags and dependency trees.

We also perform ablation studies to justify the effectiveness of the proposed components. We compare the complete method with its variants without Alignment-enhanced Training or Iterative Inference.

	WIKIEVENTS			ACE 2005		
	Train	Validation	Test	Train	Validation	Test
# Event Types	49	25	34	33	-	-
# Argument Types	57	32	44	22	-	-
# Events	3241	345	365	4202	450	403
# Sentences	5262	378	492	17172	923	832
# Documents	206	20	20	351	80	80

Table 5.3: Dataset statistics for document-level event argument extraction.

5.3.2 Implementation Details.

We implement our models with Huggingface [165]. We train each model for 4 epochs with a batch size of 4 for baselines and 2 for EA^{2}E. The model is optimized with the Adam optimizer with a learning rate of 3e-5, $\alpha = 1$ and β=0.5. We define event neighborhood as trigger distance less than 40 tokens. For inference, the maximum number of iterations is 3. For numerical consistency, all experiment results are averaged across three random runs. The hyper-parameters are selected by grid search based on model performance on development set. β is chosen from $\{0.1, 0.5, 1\}$, the trigger distance is chosen from $\{20, 40, 60, 80, 100\}$, and the learning rate is chosen in $\{3e-5, 5e-5\}$.

5.3.3 Results

Table 5.1 and Table 5.2 show that our proposed EA^{2}E consistently performs better than strong baseline methods across datasets and evaluation metrics. In general document-level methods have better performance, especially in terms of recall, because sentence-level methods are more likely to miss cross-sentence arguments.

Alignment-enhanced training brings a significant improvement over BART-Gen but comes with higher training costs since the inputs are doubled. Iterative inference brings unstable improvement. More iterations brings higher performance only to a certain range. Since the only differences among iterations are their inputs, we conclude that labeling the arguments of other events helps the model extract the arguments of the current event. The upper bound of this improvement is limited by the error propagation in the augmented context.

Qualitative Analysis. Table 5.4 presents some representative examples. BART-Gen incorrectly assigns *Tsarnaev* to the Target role, and *police* to the attacker role in the first example. It also misses the killer *brothers* in the second example and the attacker *Laden* in the third example. The second example shows the advantage of the Alignment-enhanced

Input	Dzhokhar Tsarnaev visits Silva and borrows the Ruger pistol — the gun that was later used to kill MIT police officer Sean Collier and during the **shootout** with police in Watertown.
Gold	Target: police, Instrument: gun
BART-Gen	Target: Tsarnaev, Attacker: police, Place: Watertown
EA^{2}E w/o II	Target: police, Attacker: Tsarnaev, Place: Watertown
EA^{2}E	Target: police, Attacker: Tsarnaev, Place: Watertown
Input	The brothers allegedly set off two bombs alongside the Boston Marathon course, **killing** three people and injuring 264.
Gold	Killer: brothers, Victim: people
BART-Gen	Victim: people
EA^{2}E w/o II	Killer: brothers, Victim: people
EA^{2}E	Killer: brothers, Victim: people
Input	Osama bin Laden is charged with masterminding the 1998 bombings of two U. S. embassies in East Africa, believed to have had a role in the October 2000 **attack** on the USS Cole in the Yemeni port of Aden.
Augmented Input	<tag> Attacker < \tag> Osama bin Laden is charged with masterminding the 1998 bombings of two U. S. <tag> Target < \tag> embassies in <tag> Place < \tag> East Africa, believed to have had a role in the October 2000 **attack** on the USS Cole in the Yemeni port of Aden.
Gold	Target: Cole, Target: port, Attacker: Laden
BART-Gen	Target: Cole, Place: Yemeni
EA^{2}E w/o II	Target: Cole, Place: Yemeni
EA^{2}E	Target: Cole, Target: port, Attacker: Laden

Table 5.4: Examples of extracted arguments from BART-Gen, EA^{2}E w/o II, and EA^{2}E. We label **target event mention** with bold, gold arguments with underlines, correct but not annotated arguments with waves, and incorrect arguments with red. In the third example, we also present the augmented input for Iterative Inference, in which the arguments of the bombing event are tagged.

Training component in EA^{2}E, which helps extract the killer argument. The third example shows how Iterative Inference works with the augmented input: The tagged attacker in the neighboring *bombing* event is also the attacker in the target *attack* event.

Remaining Challenges. Though effective, Iterative Inference may propagate errors among iterations. In addition, the success of event awareness relies on the assumption that events in a neighborhood defined by trigger distance are interrelated to the target event. However, this assumption is not always held true in the case that distant events bring redundant information. It is not necessarily hurting the information but it brings noise by

incorrectly implying the relations between the distant events and the target event.

5.4 CONCLUSION

We introduce Event-Aware Argument Extraction (EA^2E) model to improve self-contained consistency in document-level event argument extraction. We empirically validate the contributions of alignment-enhanced training and iterative inference. We conclude that iterative inference brings higher performance only to a certain range of iterations and alignment-enhanced training brings significant improvement with costs.

CHAPTER 6: EFFICIENT NYSTRÖM-BASED LONG TRANSFORMER

The improvements on consistency usually bring heavier methodology and efficiency problems. Large language models (LLM) are growing deeper, wider, and smarter, but have become increasingly difficult to train due to the growing computation costs. One of the computation bottlenecks is in the computation of attention layers because LLMs are mostly in the structure of transformers. Therefore, in the current era of large language models, it is particularly important to solve the efficiency problem. In the following two chapters, we first identify the criteria of efficiency and propose methods for improving computation efficiency for transformers.

6.1 CRITERIA OF EFFICIENCY

Efficiency in NLP can be categorized by resource type into parameter efficiency, training efficiency, data efficiency, and computation efficiency. Parameter efficiency focuses on reducing the number of model parameters while keeping comparable performance through model compression, pruning, distillation, lottery network discovery, adapter-based tuning, etc. Training efficiency optimizes the training strategies, such as early stopping, learning schedule, and prompt tuning, to significantly reduce training costs. Data efficiency exploits limited training samples with data augmentation, few-shot data, meta-learning, active learning, etc. Computation efficiency generally reduces computation resources by optimizing the calculation process with parameter approximation, quantization, sparsification, etc.

In this thesis proposal, we mainly focus on **computation efficiency**, whose evaluation usually involves measuring training/inference speed, CUDA memory usage, largest batch size, training power usage, etc. For example, [166] propose to evaluate the power efficiency with the total energy consumption during training, training speed, inference speed, and power efficiency, and conclude that increasing model size is more power efficient than increasing sequence length in higher accuracy.

For fair and easy comparison, in this work we consider a long transformer to be computationally efficient when it (1) reduces space complexity and supports longer sequence and larger batch size, (2) reduces time complexity with less training time per step and less total time to converge, (3) requires the same or less amount of data to reach comparable performance with vanilla softmax without much loss from approximation.

6.2 MOTIVATION

Transformers cannot support long sequence processing and large batch size with limited resources because of their computation bottlenecks in the self-attention mechanism. Another issue of Transformers is the training instability that small perturbations in parameter updates tend to be amplified, resulting in significant disturbances in the model output [167].

The inherent connection between Gaussian Kernel and Softmax operation [13] motivates us to replace the softmax structure with Gaussian kernels. Kernelized Attention empirically stabilizes the model training while being comparable to self-attention in model accuracy. However, it is not superior in computation efficiency. Therefore, we propose **Skyformer** (**S**ymmetrization of **K**ernelized attention for N**Y**ström method) to accelerate kernelized attention. Skyformer adapts the Nyström method to the non-PSD empirical Gaussian kernel matrix by instead lifting the kernelized attention score matrix into a large PSD matrix that contains the un-normalized attention score matrix as the off-diagonal block. Our experiments on the LRA benchmark show that Skyformer consistently uses less space and time while achieving better accuracy than other baseline methods.

6.3 METHOD

6.3.1 Kernelized Self-attention for Training Stability

Transformers on some NLP tasks have shown to be sensitive to hyper-parameters, learning schedulers, or even random seeds, which usually demand a time-costly grid search for the best configuration in real-world applications. It has also been observed in our experiments that a slight change in the learning rate may cause the failure of convergence for some models. As pointed out by [167], small perturbations in parameter updates tend to be amplified, resulting in significant disturbances in the model output.

We conjecture that the training instability in Transformer training comes from the softmax structure, as the un-normalized attention score matrices before softmax operation tend to have extremely large condition numbers due to its fast singular value decay. Therefore, we propose to replace the softmax structure with Gaussian kernels.

For a given input sequence $X \in R^{n \times d_0}$ of length n and embedding dimension d_0, The dot-product attention for a single head in Transformer [96] is defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{p}}\right)V \tag{6.1}$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$, and W_Q, W_K and W_V are the query, key, and value weight metrics that linearly project the input X of d_0 dimension to an output tensor of p dimensions.

Kernelized Attention replaces the softmax structure in vanilla self-attention with a Gaussian kernel, and the new attention model is stated as:

$$\text{Kernelized-Attention}(Q, K, V) = BV = \kappa\left(\frac{Q}{p^{1/4}}, \frac{K}{p^{1/4}}\right)V \tag{6.2}$$

We define the n-by-n matrix B as the kernelized attention score matrix $\kappa(Q/p^{1/4}, K/p^{1/4})$. Each element b_{ij} from the i-th row and j-th column in B is equal to $\phi(q_i, k_j)$, where q_i (resp. k_j) is the i-th (resp. j-th) row in Q (resp. K).

6.3.2 Applying Nyström Method for Training Efficiency

We apply the Nyström method that replaces B with its low-rank approximation $\tilde{B}$ to reduce calculations:

$$\tilde{B} = BS(S^T BS)^\dagger S^T B \tag{6.3}$$

, where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix, and $S \in R^{n\times d}$ is a zero-one sub-sampling matrix whose columns are a subset of the columns in I, indicating which d observations have been selected.

The adaptation of Nyström method requires B to be positive semi-definite (PSD), while B is an asymmetric (and thus non-PSD) empirical kernel matrix constructed with two different n-by-p design matrices Q and K.

To tackle the challenge of approximating a non-PSD matrix B, our first step is to complete the matrix into a PSD matrix $\bar{B}$:

$$\bar{B} = \phi(\begin{pmatrix} Q \\ K \end{pmatrix}, \begin{pmatrix} Q \\ K \end{pmatrix}) \tag{6.4}$$

Then we approximate $\bar{B}$ with $\bar{\tilde{B}}$ through

$$\bar{\tilde{B}} = \bar{B}S(S^T \bar{B}\mathbf{S})^\dagger \mathbf{S}^T \bar{B}, \tag{6.5}$$

where S is a $2n$-by-d uniform sub-sampling matrix.

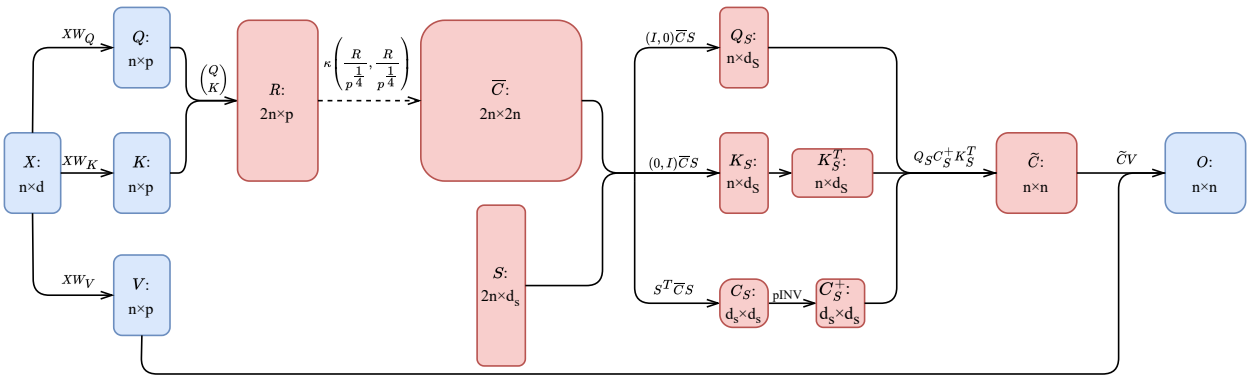

Figure 6.1: We apply the Nyström method to Kernelized Attention to reduce calculations.

The final approximation will be given as

$$\tilde{B} = (I, 0)\tilde{\tilde{B}}(0, I)^T. \tag{6.6}$$

6.4 EVALUATION

6.4.1 LRA Benchmark

We evaluate the proposed methods on five document-level classification tasks from Long Range Arena (LRA) benchmark [116], which focuses on model quality under long-context scenarios. The LRA benchmark covers diverse long-sequence tasks in sequence length, task difficulty, and inspected model abilities. We exclude Pathfider-X because it fails all baseline models.

- **ListOps** [168]: This 10-label classification task requires the models to parse a sequence of length $2k$ of numbers and operators and evaluates their capacity of modeling hierarchically structured long sequences.
- **Text Classification** on IMDb review dataset[169]: This byte-Level binary classification task requires the model to analyze the sentiment of a sequence of length $4k$ by composing the unsegmented characters into higher-level meaningful units.
- **Document Retrieval** on AAN dataset [170]: This byte-Level binary classification task requires the model to compress long sequences of length $4k$ into representations for similarity score calculation in a two-tower setup without cross-attention.
- **Pathfinder** on CIFAR-10 dataset [171]: This binary classification task requires the model to decide whether two points are connected by a dashed path on an image

represented as a pixel sequence of length $4k$, and exams their capacity to capture long-range spatial dependency.

- **Image Classification** [172]: This 10-label classification task requires the models to learn the spatial relations between the flattened input pixels of length $1k$.

We report the classification accuracy on the test set, training time, and peak memory usage during training for each task.

6.4.2 Baselines

As it is not realistic to exhaustively fine-tune all models and search for the best performance under limited computation resources, we instead only replace the self-attention module with the various attention methods and keep other experimental settings the same for fair comparisons.

Aside from the vanilla quadratic self-attention, we compare with Big Bird [104], Performer [13], Linformer [108], Nyströmformer [112], Informer [105], and Reformer [106]. Most methods are approximating the vanilla full attention for efficiency and thus are not expected to have better performance.

For clarification, deep transformers or pretrained language models are not appropriate baselines. Training a deep transformer from scratch requires large computational resources and much more data to converge, and therefore is not adopted by previous work. A shallow transformer structure, on the other hand, has been justified by previous work to be enough for a fair comparison in attention acceleration performance. Pretrained models are trained for token-level text-based tasks and are not suitable for image pixel sequences (as in Pathfinder and Image Classification), character sequences (as in Text Classification) and math operation sequences (as in ListOps).

6.4.3 Implementation Details

We conduct each experiment on a single Tesla V100 SXM2 16GB. We use the LRA evaluation benchmark reimplemented in PyTorch. Following [112] we use a 2-layer transformer model with 64 embedding dimension, 128 hidden dimension, 2 attention heads, and mean pooling for classification. Batch size is selected conditioned on the memory requirements of the standard self-attention method, which leads to 16 for Text Classification, 32 for ListOps, 16 for Document Retrieval, 128 for Pathfinder, and 256 for Image Classification. Learning rate is set to $1e-4$ for Text Classification, ListOps, and Image Classification, and $2e-4$

Model	Text	ListOps	Retrieval	Pathfinder	Image	AVG.
Self-Attention	61.95	38.37	80.69	65.26	40.57	57.37
Kernelized Attention	60.22	38.78	81.77	70.73	41.29	58.56
Nystromformer	64.83	38.51	80.52	69.48	41.30	58.93
Linformer	58.93	37.45	78.19	60.93	37.96	54.69
Informer	62.64	32.53	77.57	57.83	38.10	53.73
Performer	64.19	38.02	80.04	66.30	41.43	58.00
Reformer	62.93	37.68	78.99	66.49	48.87	58.99
BigBird	63.86	39.25	80.28	68.72	43.16	59.05
Skyformer	64.70	38.69	82.06	70.73	40.77	**59.39**

Table 6.1: Classification accuracy (%) on LRA benchmark in fixed-step setting.

for Retrieval and Pathfinder. For numerical consistency, all experiment results are averaged across three runs with different random seeds.

For comparable computation complexity, we control the number of features used in all methods, which leads to 256 as the number of features in Skeinformer, 256 as k in Linformer, 256 as the number of landmarks in Nyströmformer, $(256/\log n)$ as the factor in Informer, and 256 as the number of features in Performer. Additionally, the number of random blocks and block size in Big Bird are by default 3 and 64, under which setting Big Bird will visit $640 \cdot n$ elements in the attention matrix while other models visit $256 \cdot n$ elements.

6.4.4 Results

Each model on each task is trained for $50k$ steps, during which the best checkpoint with the highest accuracy on the development set will be saved for evaluation.

We do not follow all settings in [112] due to hardware limitations. The compromises, such as approximation dimension and gradient accumulations steps, might bring performance differences compared to results reported in [112]. The training instability problem also helps explain the performance gap.

The training process of the standard softmax-based method is unstable: it takes more steps to reach the stationary distribution of its long-time limit, and it is more easily getting stuck in a local minimum. Runs with different random seeds may bring divergent performances, and probably leads to lower averaged scores. We have also tried directly approximating the self-attention method with the Nyström method and observed numerical instability during training.

Replacing the softmax structure with Gaussian kernel somehow alleviates this instability

Model	Time (h)					Memory (GB)				
	TC	LO	RE	PF	IC	TC	LO	RE	PF	IC
Self-Attention	4.30	2.24	8.33	2.57	4.22	10.37	5.37	10.77	5.74	11.47
KA	3.91	1.99	7.46	2.42	4.05	5.73	5.94	10.46	6.38	6.38
Nystromformer	0.71	0.71	1.29	1.49	2.70	1.21	1.37	2.39	3.35	6.71
Linformer	0.65	0.60	1.13	1.09	2.19	0.99	0.99	1.89	1.97	3.94
Informer	1.60	1.19	2.91	2.39	3.90	5.12	4.85	5.77	4.75	9.51
Performer	0.77	0.73	1.41	1.40	2.55	1.09	1.09	2.16	2.20	4.39
Reformer	0.94	0.85	1.73	1.70	3.08	1.61	1.61	2.98	3.21	6.42
BigBird	2.00	1.88	3.81	3.39	6.53	2.83	2.71	4.97	4.97	9.95
Skyformer	1.02	1.29	1.86	2.03	3.40	1.59	1.75	3.15	4.13	8.26

Table 6.2: Running time (hour) and peak memory usage (GB) in fixed-step setting. **TC**: Text Classification. **LO**: ListOps. **RE**: Retrieval. **PF**: Pathfinder. **IC**: Image Classification. **KA**: Kernelized Attention.

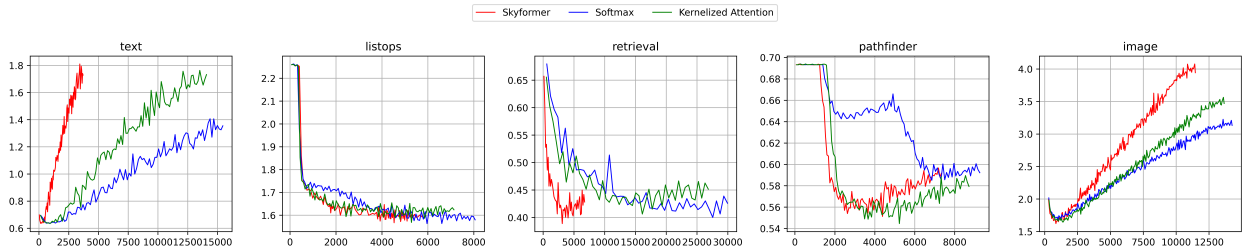

Figure 6.2: Validation loss changes for $50k$ steps. X-axis: Training time (second). Y-axis: Cross Entropy Loss on the validation set.

problem with boosted performance as shown in Table 6.1. However, the time and space requirement of Kernelized Attention is not significantly improved compared to the original version, which serves as the motivation to approximate Kernelized Self-Attention with Nyström method.

Though not necessarily the fastest, our proposed Skyformer can efficiently converge to the long-time limit with comparable general performance in classification accuracy (Table 6.1) and resource consumption (Table 6.2). The advantages over standard self-attention are significant with consistently less training time and generally better performance. For example, Skyformer brings nearly 4 times speed-up on text classification and document retrieval while with 2.75% and 1.37% accuracy improvement over the standard self-attention.

Figure 6.2 shows the validation loss changes with respect to training time for 50k steps as supplementary results for the experiments above. In general, Skyformer converges faster and finishes 50k steps earlier than vanilla Attention and Kernelized Attention over all tasks.

We further remark that on Text Classification, all models quickly fall into over-fitting, and thus the validation losses rise quickly. On Pathfinder, due to the difficulty of training, in the trial shown in the figure vanilla Attention fails to reach the best long-time limit under a certain setting.

6.5 CONCLUSION

Motivated by the connection between kernel methods and self-attention, we introduce Kernelized Attention, which replaces the softmax structure in self-attention with a Gaussian kernel. We also propose Skyformer, which adapts the Nyström method to Kernelized Attention to improve its efficiency. We expect the new model can enjoy more stable training while inheriting the strong performance from self-attention. Extensive experiments verify our intuitions and show that both Kernelized Attention and its Nyström approximation variant have comparable accuracy to the original Transformer on the LRA benchmark.

The direct development of this work is the incorporation of further computation tricks in kernel methods. Other related questions include the choice of the kernel other than the Gaussian kernel in our kernelized attention model. It is expected that for different tasks there will be specific kernels more proper than the original self-attention. The results of this work also shed new light on the design of the attention mechanism, which may benefit board downstream NLP tasks.

CHAPTER 7: EFFICIENT SKETCHING-BASED LONG TRANSFORMER

The last chapter proposes to replace the softmax operation with a kernel to improved training stability. In this chapter, we keep the softmax structure and turns to the sketching framework for improving efficiency.

7.1 MOTIVATION

Among the efficient transformer methods introduced in Section 2.4, Linformer [108] and Informer [105] are two representative approaches to reducing the quadratic self-attention to a linear operation in both space and time complexity. Linformer forms a low-rank factorization of the original attention by decomposing it into smaller attentions, while Informer allows each key to only attend to a certain number of queries.

To better understand Linformer and Informer, we introduce the sketching framework [173] for self-attention, where sketching methods replace the original matrix B with its random sketch BS to reduce computations. For the perspective of matrix approximation, we found that Informer and Linformer either do not fully utilize the information in the value matrix, or deviate from the original self-attention output.

These observations motivate us to introduce Skeinformer, which applies sub-sampling sketching to reduce time complexity. Skeinformer exploits the information from the value matrix V with column sampling and incorporates an adaptive row normalization step, which approximates the unselected rows by a vector with all elements as 1 over n. In addition, the pilot sampling reutilization step reuses the computation from pilot sampling to improve both approximation accuracy and training efficiency.

7.2 METHOD

7.2.1 Background: Sketching Method

Sketching methods replace the original matrix B with its random sketch BS to reduce computations. In practice, to apply the sketching method we plug an identity matrix into the original expression, and then formally replace the identity matrix with the product SS^T, as the distribution of S is usually designed to satisfy the constraint that

$$\mathcal{E}(SS^T) = I. \tag{7.1}$$

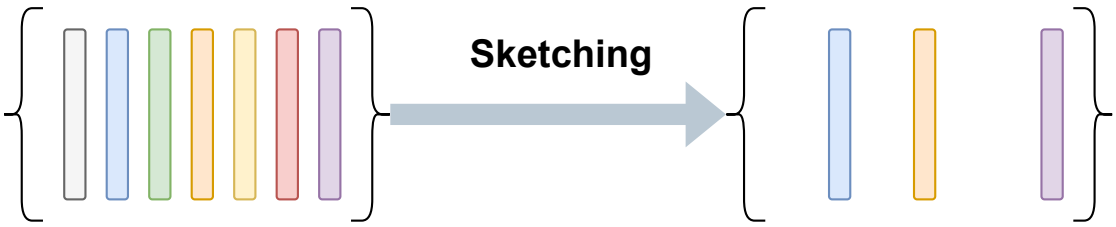

Figure 7.1: Sketching methods replace the original matrix B with its random sketch BS to reduce computations.

Common methods to construct a sketching matrix include sub-Gaussian maps [174, 175], subsampled randomized Hadamard transform [176, 177, 178], sparse oblivious subspace embeddings [179], very sparse random projection [180], and sub-sampling sketching [181]. Specifically, Informer and Linformer, two efficient transformer-based methods, can be understood as applications of sub-sampling sketching and sub-Gaussian maps, respectively.

7.2.2 Self-attention Approximation

Self-attention can be written as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{p}}\right)V = D^{-1}AV \tag{7.2}$$

where $A = \exp(\mathbf{QK}^T/\sqrt{p})$, and D is a diagonal matrix whose diagonal is $\exp(QK^T/\sqrt{p}) \cdot 1$ (1 is a size-n vector with all elements being 1).

A naïve step in applying sketching method to approximate the self-attention output $D^{-1}AV$ is to construct a random sketch of the un-normalized attention score matrix A, the bottleneck in computation.

From this perspective, Informer and Linformer construct two types of sketches, A^TS and AS respectively. Informer selects d important rows of $D^{-1}A$ high sparsity measurement to represent $D^{-1}A$. This process can be related to a sketched approximation $D^{-1}SS^TA$, where S is a sub-sampling matrix. Another type of sketch AS is mentioned (but not finally used) in Linformer. To avoid the computation of the whole matrix A, Linformer replaces the form of sketching method with $\text{softmax}\left((QK^T/\sqrt{p})S\right)S^TV$, which sacrifices the accuracy for efficiency in some tasks as shown in later experimental results.

Algorithm 7.1: Skeinformer

Input: query matrix Q, key matrix K, value matrix V (all are n-by-p), and sub-sample size d

Output: Attention output matrix R with the same shape as V

Uniformly sample d indices $j_1, \cdots, j_d$ with replacement;

Construct the $d \times p$ matrix Q_J as to the index set $J = \{j_k\}_{k=1}^d$, whose k-th row is $Q_{(j_k)}$;

Compute the matrix $B_J = \text{softmax}\left(Q_J K^T/\sqrt{p}\right)$; `// pilot sampling`

Based on B_J, give the estimated sub-sampling probabilities $\{\hat{p}_i\}_{i=1}^n$;

With $\{\hat{p}_i\}_{i=1}^n$ sample d indices $j'_1, \cdots, j'_d$ without replacement;

Construct the d-by-p matrix $K_{J'}$ (resp., $V_{J'}$) according to the indices list $J' = \{j'_k\}_{k=1}^d$, whose k-th row is $K_{(j'_k)}$ (resp., $V_{(j'_k)}$);

Compute the two matrices $A^{J'} = \exp\left(QK_{J'}^T/\sqrt{p}\right)$, and $R_{J'} = A^{J'}V_{J'}$; `// column sampling`

Construct a length n column vector g whose i-th element is $(\prod_{k=1}^d a_{ij'_k})^{\frac{1}{d}}, \forall i \in [n]$;

Compute the row sum vector $d = A^{J'}1_d + (n-d)g$; `// adaptive row normalization`

Denote the un-selected part of V as $V_{(J')^C}$, and compute the vector $V = V_{(J')^C}^T 1_{n-d}$;

Obtain the intermediate output $R = \text{diag}(d^{-1})(R_{J'} + gV^T)$, where d^{-1} is the element-wise inverse of d;

Compute $B_J V$ and assign it to the corresponding rows of R ; `// pilot sampling reutilization`

Return the matrix R as the ultimate output of this algorithm;

7.2.3 Skeinformer

Skeinformer consists of three components: the initial column sampling that incorporates the information from the value matrix V into the sampling probabilities, the adaptive row normalization that fills un-selected columns with the averaged selected columns and the pilot sampling re-utilization. We describe the proposed method in Algorithm 7.1.

Column Sampling. Skeinformer applies sub-sampling sketching to reduce time complexity and exploits the information from the value matrix V with column sampling.

The row selection in Informer can be further improved by utilizing the information from V:

$$D^{-1}ASS^TV, \tag{7.3}$$

where S above is a sub-sampling matrix with sampling probabilities

$$p_i \propto \|(D^{-1}A)^{(i)}\|_2 \|V_{(i)}\|_2, \quad i = 1, 2, \ldots, n. \tag{7.4}$$

We remark that using the sub-sampling sketching in this way can both circumvent the computation burden of Gaussian sketching, and also allow the incorporation of the information from V. As S formally samples some columns from $D^{-1}A$, we name the procedure column sampling in our method.

Adaptive Row Normalization. Skeinformer also incorporates an adaptive row normalization step, which fills un-selected columns with the averaged selected columns. From the model training perspective, it allows the whole value matrix V in Skeinformer to participate in the computation and thus can improve the efficiency of updating W_V during the training.

Specifically, in adaptive row normalization any row in the matrix A can be divided into two parts, the exactly computed elements in the selected columns with indices $\{j'_k\}_{k=1}^{d} \subset [n]$ and the other elements in the un-selected columns. For the latter, in each row, we set all the un-selected elements as the geometric mean of the selected ones, considering the exponentiation in softmax. We then perform row normalization based on the above construction, in which the i-th diagonal element in D is estimated as

$$\hat{d}_{ii} = \sum_{k=1}^{d} a_{ij_k} + (n-d)(\prod_{k=1}^{d} a_{ij_k})^{\frac{1}{d}}, \tag{7.5}$$

where each a_{ij} is the corresponding element in matrix A. Next, we normalize rows composed of exact elements in the selected columns and the other elements estimated with the mean value above.

Pilot Sampling Reutilization. We also introduce a simple yet effective step, pilot sampling reutilization, which reuses the computation from pilot sampling to improve both approximation accuracy and training efficiency. Since we have already computed B_J in the pilot sampling step, we can exactly reproduce the d rows in the original self-attention output with an additional product $B_J V$ in $O(n \log n)$ time. This allows for more precise approximation with little cost. In addition, the computation of those rows involves the whole key matrix K, which benefits the training of the parameters W_K.

7.3 EVALUATION

7.3.1 Implementation Details

As it is not realistic to exhaustively fine-tune all models and search for the best performance under limited computation resources, we instead replace the self-attention module in transformer with the various drop-in attention methods and keep other experimental settings

Models	Text	ListOps	Retrieval	Pathfinder	Image	Average
Standard [96]	57.69	38.15	80.10	73.59	37.97	57.50
w/o dropout	59.44	38.17	79.35	72.35	37.58	57.38
V-Mean	65.29	28.78	80.49	61.01	34.33	53.98
BigBird [104]	61.91	38.86	79.73	71.75	35.00	57.45
Performer [13]	57.67	37.70	75.69	56.50	37.40	52.99
Nystromformer [112]	60.91	37.76	79.87	72.53	31.93	56.60
Reformer [106]	62.69	37.94	78.85	69.21	36.42	57.02
Linformer [108]	58.52	37.97	77.40	55.57	37.48	53.39
w/ uJLT	59.12	37.48	79.39	68.45	35.96	56.08
Informer [105]	61.55	38.43	80.88	59.34	36.55	55.35
· w/ mask	60.98	37.26	79.92	62.51	37.19	55.57
Skeinformer	62.47	38.73	80.42	71.51	37.27	**58.08**
w/o CS	64.48	30.02	80.57	64.35	36.97	55.28
w/o RN	60.67	37.69	78.67	66.35	37.06	56.09
w/ SRN	60.26	38.35	78.97	65.41	39.72	56.54
w/o PSR	62.39	38.12	79.88	71.53	37.20	57.83

Table 7.1: Classification accuracy (%) on the test sets of LRA benchmark in flexible-step setting. The approximation methods are not expected to outperform the original methods (standard self-attention) though they surprisingly do. **CS**: Column Sampling. **uJLT**: unreduced JLT. **RN**: Row Normalization. **SRN**: Simple Row Normalization. **PSR**: Pilot Sampling Reutilization.

the same. Following [112] we use a 2-layer transformer model with 64 embedding dimensions, 128 hidden dimensions, and 2 attention heads for all experiments. Mean pooling is used in all classifiers.

For comparable computation complexity, we control the number of features used in all methods, which leads to 256 as the number of features in Skeinformer, 256 as k in Linformer, 256 as the number of landmarks in Nyströmformer, $(256/\log n)$ as the factor in Informer, and 256 as the number of features in Performer. Additionally, the number of random blocks and block size in Big Bird are by default 3 and 64, under which setting Big Bird will visit $640 \cdot n$ elements in the attention matrix while other models visit $256 \cdot n$ elements. A clearer complexity evaluation on the FLOPs of each method is provided in Appendix.

We use Adam optimizer [182] with a learning rate of $1e-4$. Batch size is selected conditioned on the memory requirements of Skeinformer, which leads to 128 for Text Classification, 256 for ListOps, 64 for Document Retrieval, 512 for Pathfinder and 256 for Image. For methods reporting out-of-memory errors, we apply gradient accumulation and report the accumulated steps. We conduct all experiments on one Tesla V100 SXM2 16GB. For numerical consistency, all experiment results are averaged across three random runs.

Models	Text		ListOps		Retrieval		Pathfinder		Image	
	time ↓	bz ↑	time ↓	bz ↑	time ↓	bz ↑	time ↓	bz ↑	time ↓	bz ↑
Standard	50.63	16	22.30	64	53.27	16	13.91	128	21.40	64
w/o dropout	39.49	8	19.50	32	41.88	8	11.79	64	14.88	32
V-Mean	3.62	128	4.14	256	3.90	64	3.67	512	4.44	256
BigBird	20.59	64	17.28	128	21.73	32	17.83	256	18.84	256
Performer	2.63	64	9.31	128	12.50	32	10.40	256	8.94	256
Nyströmformer	12.18	64	12.28	128	13.35	32	19.58	128	10.30	256
Reformer	10.53	64	8.28	128	11.27	64	9.25	256	11.88	256
Linformer	7.91	64	6.25	128	8.08	64	6.90	256	6.65	256
w/ uJLT	36.87	8	21.49	32	35.93	4	15.17	128	22.03	128
Informer	33.13	16	21.89	32	36.52	16	26.14	64	24.92	128
w/ mask	25.94	32	21.50	64	35.95	32	15.79	128	22.58	128
Skeinformer	9.60	64	9.66	128	10.61	64	9.25	256	11.86	256
w/o CS	7.60	128	6.66	256	6.70	64	7.27	512	7.76	256
w/o RN	25.02	16	16.02	64	55.72	4	11.12	256	15.52	128
w/ SRN	6.80	128	8.16	256	8.03	64	6.84	512	11.27	256
w/o PSR	7.15	128	7.31	256	8.68	64	7.09	512	10.19	256

Table 7.2: Training time (minute per thousand steps) and actual batch size (in batch accumulation) on LRA benchmark in the flexible-step setting. Less training time per thousand steps indicates higher time efficiency. Higher batch size indicates higher space efficiency, and within a certain range means more accurate gradient estimations. **CS**: Column Sampling. **uJLT**: unreduced JLT. **RN**: Row Normalization. **SRN**: Simple Row Normalization. **PSR**: Pilot Sampling Reutilization.

Instead of setting a fixed epoch number, we train all models until convergence with a stopping strategy (if better performance is not observed for 10 checking steps on the validation set we will stop the training process). we simulate the case of real-world applications of efficient transformers that models are trained with their maximum batch size conditioned on memory.

7.3.2 Results

The experiments are conducted on the LRA dataset introduced in Section 6.4.1. We conclude the results in Table 7.1 and Table 7.2 with the following observations:

Most $O(n)$ attention acceleration methods have comparable or better performance with vanilla attention. After the models converge to their long-time limits, Linformer tends to have worse performance possibly due to the violation of the sketching form, while Skeinformer has the best overall performance.

While surprising, those approximation methods tend to outperform the original trans-

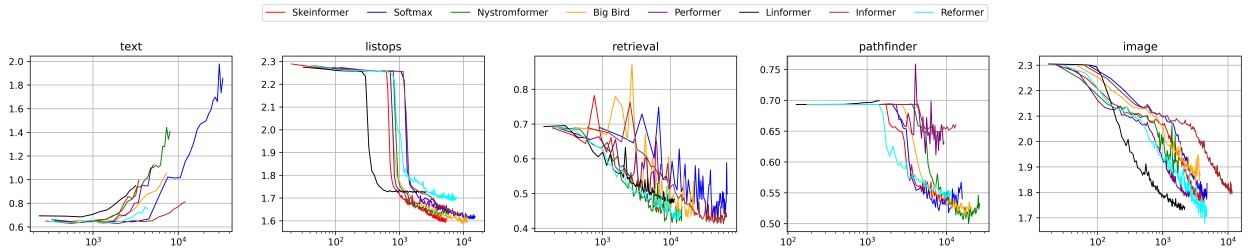

Figure 7.2: Validation loss (Y-axis) changes with regard to training time (second, X-axis).

former in most tasks. We speculate the reason behind this phenomenon is that a good approximation can recover the main signals in the original self-attention matrix, and also restrain the noise via the sparse / low-rank structure.

Skeinformer has comparable general performance in terms of time/space complexity and classification accuracy. For convergence efficiency, Skeinformer efficiently converges to the long-time limit. Regarding the training efficiency, we focus on how soon the model can attain the stationary distribution of its long-time limit [183]. The loss decay plot on ListOps in Appendix shows significant differences in the convergence rate of each method in addition to classification accuracy.

Though our method does not always outperform others (with the fastest convergence or the highest accuracy), but we remark that Skeinformer attains the best accuracy-efficiency trade-off based on experimental results. On the opposite, some model converges fast but gets stuck in a local optimum, like Linformer in some cases.

7.3.3 Validation Loss

We present the loss decay plots on all tasks in Figure 7.2. In the first subplot for the text classification task, we note all the methods quickly overfit the dataset. In all the other plots, our methods show the ability to both efficiently converge to the long-time limit and find better local minima with lower validation loss.

7.3.4 Hyper-parameter Sensitivity

Figure 7.3 shows the accuracy and training time for Skeinformer using different batch sizes (64,128) and learning rates $(1e-3, 1e-4, 1e-5)$ on text classification. The results are averaged across random trials. A larger batch size is not supported by the CUDA memory limitation. We observe that a smaller learning rate offers slower convergence but to a better point.

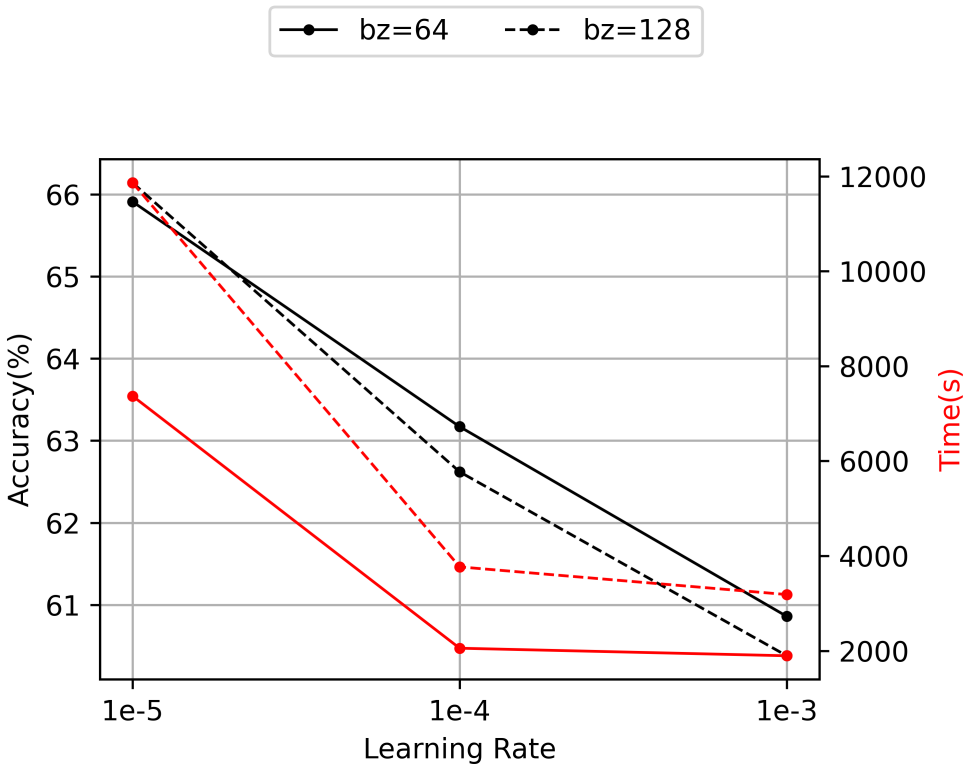

Figure 7.3: Testing accuracy and training time (Y-axis) change with regard to learning rate (X-axis).

7.4 CONCLUSION

In this chapter, we introduce and empirically validate the contributions of the components in Skeinformer, including column sampling, adaptive row normalization, and pilot sampling reutilization, with extensive comparisons with various baseline and ablation methods.

Direct further development directions include

- Applying long transformers to more long-form downstream tasks with consideration of consistency, coherence, robustness, interpretability, etc.
- Extending self-attention acceleration methods to encoder-decoder cross-attention and cross-modality attentions and inspecting the performance consistency across attention settings.
- Inspecting the reason why some approximation methods tend to outperform the original transformer in most tasks in terms of accuracy.

CHAPTER 8: CONCLUSIONS, LIMITATIONS, AND FUTURE DIRECTIONS

8.1 CONCLUSIONS

Although existing Pretrained Language Models (PLMs) and Large Language Models (LLMs) have achieved significant success in document understanding, understanding long documents remains a challenge because they cannot handle consistency and efficiency problems unique to long documents. In this thesis, we study consistency and efficiency improvement methods in four typical scenarios in Long Document Understanding:

- **Consistent meta-review generation**: The generated meta-reviews should be consistent with the comments from individual reviews and the final decision. To achieve this goal, we first benchmark the task of scientific opinion summarization by collecting paper, then propose a checklist-guided iterative self-refinement approach that guides the LLM to generate consistent content, and finally construct an evaluation framework to comprehensively evaluate the quality of the generated meta-reviews.
- **Consistent document summarization**: Current abstractive summarization models often generate inconsistent content, i.e. texts that are not directly inferable from the source document, are not consistent with respect to world knowledge, or are self-contradictory. To improve the general consistency we introduce EnergySum, where we apply the Residual Energy-based Model by designing energy scorers that reflect each type of consistency and incorporating them into the sampling process.
- **Consistent event argument extraction**: Events are connected. Recent work on document-level event argument extraction models each individual event in isolation and therefore causes inconsistency among extracted arguments across events, which will further cause discrepancies for downstream applications. To address this problem, we formulate event argument consistency as the constraints from event-event relations under the document-level setting and introduce the Event-Aware Argument Extraction (EA^2E) model with augmented context for training and inference.
- **Efficient long sequence processing**: Transformer-based models are inefficient in processing long sequences due to the self-attention modules' quadratic space and time complexity. To address this limitation, we introduce two methods for self-attention acceleration, a modified Nyström method (Skyformer) to accelerate kernelized attention and stabilize training and a Sketching-based method (Skeinformer) applying subsampling sketching to accelerate self-attention.

8.2 LIMITATIONS

Throughout our studies, we have investigated several consistency and efficiency improvement methods and validated their effectiveness with extensive experiments at the time of publishing. However, most of the methods in this thesis are proposed prior to the emergence of Large Language Models and only used models of much smaller scale of parameters, which means their performance may not be steadily competitive today. Nevertheless, they are still meaningful as they have provided principled and scalable solutions for certain applications.

In addition, some methods are evaluated only in limited experiment settings because of computation limitations. For example, Skyformer and Skeinformer are evaluated with 2-layer transformers, and their effectiveness when applied to larger models remains unknown. Scaling large transformers to a long context is an important direction for LLM [184]. Meta-review generation is only evaluated with GPT-3, while the performances on Large Language Models of other sizes or model families are unexplored.

8.3 FUTURE WORK

8.3.1 Consistency

One major future research direction is to inspect the ability to capture long-range dependency for Large Language Models. Though LLM can process a longer context window in the input document, it is shown that LLM is easily distracted by irrelevant information [185]. In our experiments of meta-review generation, we have also observed that when the given context is long, the model is likely to get lost and ignore some of the inputs. We conclude this phenomenon as attention dilution. A primary solution to solve this problem without major changes to the LLM is to perform an extract-then-process pipeline, which is inspired by the classical extract-then-summarize method in summarization. Similarly, the LLM will be instructed to provide a summary of the input first, either extractive or abstractive, then perform its actual task, such as discovering the relation between two events in the document. By summarizing the given input as the first step, the LLM can have a shorter and more condensed window to perform necessary operations. With extract-then-process, long-range dependency across the input can be captured and more clearly focused on.

Another important direction to look into is hallucination mitigation. Large Language Models are known to easily hallucinate [186], especially when the given text control is loose. One solution we have tried in this thesis is to ask the model to provide evidence from the source document when giving opinions. This method is simple but effective. To further

provide textual control for mitigating hallucination, it is worth exploring incorporating the role of the verifier into the LLM itself. If an LLM has the ability to identify the source of an opinion, then it is expected to classify the source of an opinion into the input or the retrieved memory from training. If the source is not reliable, the opinion is likely to be hallucinated. Therefore, the dual role of the generator and verifier of LLM is worth exploring.

8.3.2 Efficiency

The efficient methods for transformers proposed in this thesis are designed for and experimented with encoder self-attention. The methods can naturally be applied to cross-attentions, such as encoder-decoder attention and cross-modality attention. The application can be optimized and customized with consideration of actual applications. For example, in the case when one modality has a smaller length and smaller vocabulary, we may customize the efficient attention mechanism with a caching module to improve space utilization and reduce memory costs.

Another interesting topic in this direction is to inspect the reason why some approximation methods tend to outperform the original transformer in most tasks in terms of accuracy if well-trained. A hypothesis is that some part of the input is of little importance for the understanding process and its existence will in fact act as a noise. The process of approximation is the process of denoising. Certain approximation methods can effectively separate the signals from the noises by identifying and disregarding the less important parts of the input, thus allowing the model to focus on the more meaningful components. Also, the approximation methods inherently incorporate some form of regularization, reducing over-fitting, and leading to better generalization performance on unseen data.

8.3.3 Challenges other than Consistency and Efficiency

In addition to consistency and efficiency, there are broader challenges in long document understanding. For example, understanding the discourse structure and the non-text structure in a long document is necessary for some applications such as financial annual reports. In a typical long document understanding system, the document is sequentialized to be fed as input, which will cost the loss of its structure information and possibly some of the non-text information.

Another challenge comes from robustness. When the given input is relatively long, inserting backdoor attacks into the text is imperceptible. Therefore, the model may be vulnerable to textual attack.

8.3.4 Applications

The long document understanding systems have broad applications across various areas.

First, long document understanding systems are very important for retrieval-augmented and knowledge-augmented tasks, which often involve multi-source and lengthy inputs in the form of scientific literature, knowledge base, etc. By efficiently understanding these long documents, the system can retrieve relevant information, uncover hidden correlations, and provide insights for the decision-making processes.

Second, long document understanding systems can support healthcare, legal, and financial document processing. The language used in such documents often includes domain-specific terminologies that may be challenging for people to comprehend. Long document understanding systems can extract key pieces of information, translate the terminologies, and present them in a more laymen friendly form, thereby enhancing accessibility and understanding for non-experts.

Third, long document understanding techniques directly enable an integrated Artificial Intelligent (AI) Assistant to understand and respond effectively to complex queries by processing extensive information in long documents.

REFERENCES

[1] G. Bao, Y. Zhang, Z. Teng, B. Chen, and W. Luo, "G-transformer for document-level machine translation," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.267 pp. 3442–3455.

[2] L. Zhang, T. Zhang, H. Zhang, B. Yang, W. Ye, and S. Zhang, "Multi-hop transformer for document-level machine translation," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-main.309 pp. 3953–3963.

[3] X. Tan, L. Zhang, D. Xiong, and G. Zhou, "Hierarchical modeling of global context for document-level neural machine translation," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/D19-1168 pp. 1576–1585.

[4] X. Kang, Y. Zhao, J. Zhang, and C. Zong, "Dynamic context selection for document-level neural machine translation via reinforcement learning," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.175 pp. 2242–2254.

[5] L. Huang, S. Cao, N. N. Parulian, H. Ji, and L. Wang, "Efficient attentions for long document summarization," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-main.112 pp. 1419–1436.

[6] R. Dangovski, M. Shen, D. Byrd, L. Jing, D. Tsvetkova, P. Nakov, and M. Soljacic, "We can explain your research in layman's terms: Towards automating science journalism at scale," in Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17507 pp. 12 728–12 737.

[7] S. Wang, L. Zhou, Z. Gan, Y. Chen, Y. Fang, S. Sun, Y. Cheng, and J. Liu, "Cluster-former: Clustering-based sparse transformer for long-range dependency encoding," CoRR, vol. abs/2009.06097, 2020. [Online]. Available: https://arxiv.org/abs/2009.06097

[8] R. Jia, C. Wong, and H. Poon, "Document-level n-ary relation extraction with multiscale representation learning," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/n19-1370 pp. 3693–3704.

[9] G. Nan, Z. Guo, I. Sekulic, and W. Lu, "Reasoning with latent structure refinement for document-level relation extraction," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.141 pp. 1546–1557.

[10] S. Zeng, R. Xu, B. Chang, and L. Li, "Double graph based reasoning for document-level relation extraction," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.127 pp. 1630–1640.

[11] K. Huang, S. Tang, and N. Peng, "Document-level entity-based extraction as template generation," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.emnlp-main.426 pp. 5257–5269.

[12] S. Zhang, C. Wong, N. Usuyama, S. Jain, T. Naumann, and H. Poon, "Modular self-supervision for document-level relation extraction," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.emnlp-main.429 pp. 5291–5302.

[13] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller, "Rethinking attention with performers," CoRR, vol. abs/2009.14794, 2020. [Online]. Available: https://arxiv.org/abs/2009.14794

[14] Y. Chen, Q. Zeng, H. Ji, and Y. Yang, "Skyformer: Remodel self-attention with gaussian kernel and nyström method," CoRR, vol. abs/2111.00035, 2021. [Online]. Available: https://arxiv.org/abs/2111.00035

[15] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," CoRR, vol. abs/2004.05150, 2020. [Online]. Available: https://arxiv.org/abs/2004.05150

[16] S. Ebner, P. Xia, R. Culkin, K. Rawlins, and B. V. Durme, "Multi-sentence argument linking," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.718 pp. 8057–8077.

[17] S. Zheng, W. Cao, W. Xu, and J. Bian, "Doc2edag: An end-to-end document-level framework for chinese financial event extraction," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/D19-1032 pp. 337–346.

[18] Q. Zeng, M. Sidhu, H. P. Chan, L. Wang, and H. Ji, "Meta-review generation with checklist-guided iterative introspection," CoRR, vol. abs/2305.14647, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.14647

[19] Q. Zeng, Q. Yin, Z. Li, Y. Gao, S. Nag, Z. Wang, B. Yin, H. Ji, and C. Zhang, "Improving consistency for text summarization with energy functions," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 Findings, Singapore, December 6-10, 2023. Association for Computational Linguistics, 2023.

[20] Q. Zeng, Q. Zhan, and H. Ji, "Ea^{2}e: Improving consistency with event awareness for document-level argument extraction," in Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.findings-naacl.202 pp. 2649–2655.

[21] Y. Chen, Q. Zeng, D. Hakkani-Tur, D. Jin, H. Ji, and Y. Yang, "Sketching as a tool for understanding and accelerating self-attention for long sequences," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.naacl-main.381 pp. 5187–5199.

[22] F. D. Keles, P. M. Wijewardena, and C. Hegde, "On the computational complexity of self-attention," in International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore, ser. Proceedings of Machine Learning Research, S. Agrawal and F. Orabona, Eds., vol. 201. PMLR, 2023. [Online]. Available: https://proceedings.mlr.press/v201/duman-keles23a.html pp. 597–619.

[23] Z. Qin, X. Han, W. Sun, D. Li, L. Kong, N. Barnes, and Y. Zhong, "The devil in linear transformer," CoRR, vol. abs/2210.10340, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2210.10340

[24] N. Sevim, E. O. Özyedek, F. Sahinuç, and A. Koç, "Fast-fnet: Accelerating transformer encoder models via efficient fourier layers," CoRR, vol. abs/2209.12816, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2209.12816

[25] Z. Qin, X. Han, W. Sun, B. He, D. Li, D. Li, Y. Dai, L. Kong, and Y. Zhong, "Toeplitz neural network for sequence modeling," 2023.

[26] J. Zhang, S. Jiang, J. Feng, L. Zheng, and L. Kong, "CAB: comprehensive attention benchmarking on long sequence modeling," CoRR, vol. abs/2210.07661, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2210.07661

[27] D. Soydaner, "Attention mechanism in neural networks: where it comes and where it goes," Neural Comput. Appl., vol. 34, no. 16, pp. 13 371–13 385, 2022. [Online]. Available: https://doi.org/10.1007/s00521-022-07366-3

[28] C. Nguyen, H. Man, and T. Nguyen, "Contextualized soft prompts for extraction of event arguments," in Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics, July 2023. [Online]. Available: https://aclanthology.org/2023.findings-acl.266 pp. 4352–4361.

[29] W. Liu, S. Cheng, D. Zeng, and Q. Hong, "Enhancing document-level event argument extraction with contextual clues and role relevance," in Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics, July 2023. [Online]. Available: https://aclanthology.org/2023.findings-acl.817 pp. 12 908–12 922.

[30] Y. Yang, Q. Guo, X. Hu, Y. Zhang, X. Qiu, and Z. Zhang, "An AMR-based link prediction approach for document-level event argument extraction," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, July 2023. [Online]. Available: https://aclanthology.org/2023.acl-long.720 pp. 12 876–12 889.

[31] M. Li, A. Zareian, Q. Zeng, S. Whitehead, D. Lu, H. Ji, and S. Chang, "Cross-media structured common space for multimedia event extraction," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.230 pp. 2557–2568.

[32] M. Li, Q. Zeng, Y. Lin, K. Cho, H. Ji, J. May, N. Chambers, and C. R. Voss, "Connecting the dots: Event graph schema induction with path language modeling," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.50 pp. 684–695.

[33] Q. Wang, Q. Zeng, L. Huang, K. Knight, H. Ji, and N. F. Rajani, "Reviewrobot: Explainable paper review generation based on knowledge synthesis," in Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020, B. Davis, Y. Graham, J. D. Kelleher, and Y. Sripada, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.inlg-1.44/ pp. 384–397.

[34] M. Li, Y. Lin, T. M. Lai, X. Pan, H. Wen, S. Li, Z. Wang, P. Yu, L. Huang, D. Lu et al., "Gaia at smkbp 2020-a dockerlized multi-media multi-lingual knowledge extraction, clustering, temporal tracking and hypothesis generation system," in Proceedings of Thirteenth Text Analysis Conference (TAC 2020), 2020.

[35] Q. Zeng, M. Li, T. Lai, H. Ji, M. Bansal, and H. Tong, "GENE: Global event network embedding," in Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15). Mexico City, Mexico: Association for Computational Linguistics, June 2021. [Online]. Available: https://aclanthology.org/2021.textgraphs-1.5 pp. 42–53.

[36] X. Du, Z. Zhang, S. Li, P. Yu, H. Wang, T. Lai, X. Lin, Z. Wang, I. Liu, B. Zhou, H. Wen, M. Li, D. Hannan, J. Lei, H. Kim, R. Dror, H. Wang, M. Regan, Q. Zeng, Q. Lyu, C. Yu, C. Edwards, X. Jin, Y. Jiao, G. Kazeminejad, Z. Wang, C. Callison-Burch, M. Bansal, C. Vondrick, J. Han, D. Roth, S.-F. Chang, M. Palmer, and H. Ji, "RESIN-11: Schema-guided event prediction for 11 newsworthy scenarios," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, July 2022. [Online]. Available: https://aclanthology.org/2022.naacl-demo.7 pp. 54–63.

[37] L. Ren, M. Sidhu, Q. Zeng, R. Gangi Reddy, H. Ji, and C. Zhai, "C-PMI: Conditional pointwise mutual information for turn-level dialogue evaluation," in Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering. Toronto, Canada: Association for Computational Linguistics, July 2023. [Online]. Available: https://aclanthology.org/2023.dialdoc-1.9 pp. 80–85.

[38] H. P. Chan, Q. Zeng, and H. Ji, "Interpretable automatic fine-grained inconsistency detection in text summarization," in Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023. [Online]. Available: https://aclanthology.org/2023.findings-acl.402 pp. 6433–6444.

[39] R. G. Reddy, Y. R. Fung, Q. Zeng, M. Li, Z. Wang, P. Sullivan et al., "Smartbook: Ai-assisted situation report generation," arXiv preprint arXiv:2303.14337, 2023.

[40] Y. Xu, Q. Zeng, and G. Singh, "Efficient reward poisoning attacks on online deep reinforcement learning," arXiv preprint arXiv:2205.14842, 2022.

[41] M. Hu and B. Liu, "Opinion extraction and summarization on the web," in Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA. AAAI Press, 2006. [Online]. Available: http://www.aaai.org/Library/AAAI/2006/aaai06-265.php pp. 1621–1624.

[42] S. Angelidis and M. Lapata, "Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, 2018. [Online]. Available: https://doi.org/10.18653/v1/d18-1403 pp. 3675–3686.

[43] Y. Suhara, X. Wang, S. Angelidis, and W. Tan, "Opiniondigest: A simple framework for opinion summarization," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.513 pp. 5789–5798.

[44] A. Brazinskas, M. Lapata, and I. Titov, "Unsupervised opinion summarization as copycat-review generation," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.461 pp. 5151–5169.

[45] E. Chu and P. J. Liu, "Meansum: A neural model for unsupervised multi-document abstractive summarization," in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019. [Online]. Available: http://proceedings.mlr.press/v97/chu19b.html pp. 1223–1232.

[46] A. Bhaskar, A. R. Fabbri, and G. Durrett, "Zero-shot opinion summarization with GPT-3," CoRR, vol. abs/2211.15914, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2211.15914

[47] R. K. Amplayo and M. Lapata, "Unsupervised opinion summarization with noising and denoising," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.175 pp. 1934–1945.

[48] R. K. Amplayo, S. Angelidis, and M. Lapata, "Unsupervised opinion summarization with content planning," in Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17481 pp. 12 489–12 497.

[49] C. Lin and E. H. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003, M. A. Hearst and M. Ostendorf, Eds. The Association for Computational Linguistics, 2003. [Online]. Available: https://aclanthology.org/N03-1020/

[50] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[51] B. Goodrich, V. Rao, P. J. Liu, and M. Saleh, "Assessing the factual accuracy of generated text," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019, A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, Eds. ACM, 2019. [Online]. Available: https://doi.org/10.1145/3292500.3330955 pp. 166–175.

[52] E. Durmus, H. He, and M. T. Diab, "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.454 pp. 5055–5070.

[53] A. Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.450 pp. 5008–5020.

[54] T. Scialom, P. Dray, S. Lamprier, B. Piwowarski, J. Staiano, A. Wang, and P. Gallinari, "Questeval: Summarization asks for fact-based evaluation," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.529 pp. 6594–6604.

[55] F. Ladhak, E. Durmus, H. He, C. Cardie, and K. R. McKeown, "Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.100 pp. 1410–1421.

[56] T. Goyal and G. Durrett, "Evaluating factuality in generation with dependency-level entailment," in Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, ser. Findings of ACL, T. Cohn, Y. He, and Y. Liu, Eds., vol. EMNLP 2020. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.findings-emnlp.322 pp. 3592–3603.

[57] T. Goyal and G. Durrett, "Annotating and modeling fine-grained factuality in summarization," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-main.114 pp. 1449–1462.

[58] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.750 pp. 9332–9346.

[59] H. Lee, K. M. Yoo, J. Park, H. Lee, and K. Jung, "Masked summarization to generate factually inconsistent summaries for improved factual consistency checking," in Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.findings-naacl.76 pp. 1019–1030.

[60] F. Nie, J. Yao, J. Wang, R. Pan, and C. Lin, "A simple recipe towards reducing hallucination in neural surface realisation," in Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, A. Korhonen, D. R. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/p19-1256 pp. 2673–2679.

[61] Z. Zhao, S. B. Cohen, and B. Webber, "Reducing quantity hallucinations in abstractive summarization," CoRR, vol. abs/2009.13312, 2020. [Online]. Available: https://arxiv.org/abs/2009.13312

[62] D. Kang and T. Hashimoto, "Improved natural language generation via loss truncation," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.66 pp. 718–731.

[63] M. Cao, Y. Dong, and J. C. K. Cheung, "Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.236 pp. 3340–3354.

[64] T. Dixit, F. Wang, and M. Chen, "Improving factuality of abstractive summarization without sacrificing summary quality," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-short.78 pp. 902–913.

[65] Z. Zhang, H. Elfardy, M. Dreyer, K. Small, H. Ji, and M. Bansal, "Enhancing multi-document summarization with cross-document graph-based information extraction," in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023. [Online]. Available: https://aclanthology.org/2023.eacl-main.124 pp. 1696–1707.

[66] S. Cao and L. Wang, "CLIFF: contrastive learning for improving faithfulness and factuality in abstractive summarization," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.532 pp. 6633–6649.

[67] M. Cao, Y. Dong, J. Wu, and J. C. K. Cheung, "Factual error correction for abstractive summarization models," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.506 pp. 6251–6258.

[68] Y. Dong, S. Wang, Z. Gan, Y. Cheng, J. C. K. Cheung, and J. Liu, "Multi-fact correction in abstractive text summarization," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.749 pp. 9320–9331.

[69] C. Zhu, W. Hinthorn, R. Xu, Q. Zeng, M. Zeng, X. Huang, and M. Jiang, "Enhancing factual consistency of abstractive summarization," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-main.58 pp. 718–733.

[70] M. Li, L. Zhang, H. Ji, and R. J. Radke, "Keep meeting summaries on topic: Abstractive multi-modal meeting summarization," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, July 2019. [Online]. Available: https://aclanthology.org/P19-1210 pp. 2190–2196.

[71] Y. R. Fung, T. Chakraborty, H. Guo, O. Rambow, S. Muresan, and H. Ji, "Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly," arXiv preprint arXiv:2210.08604, 2022.

[72] Y. Fung, C. Thomas, R. Gangi Reddy, S. Polisetty, H. Ji, S.-F. Chang, K. McKeown, M. Bansal, and A. Sil, "InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.acl-long.133 pp. 1683–1698.

[73] V. Bhatnagar, D. Kanojia, and K. Chebrolu, "Harnessing abstractive summarization for fact-checked claim detection," in Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022. [Online]. Available: https://aclanthology.org/2022.coling-1.259 pp. 2934–2945.

[74] A. Zhiyuli, Y. Chen, X. Zhang, and X. Liang, "Bookgpt: A general framework for book recommendation empowered by large language model," arXiv preprint arXiv:2305.15673, 2023.

[75] Y. Wang, H. Le, A. D. Gotmare, N. D. Bui, J. Li, and S. C. Hoi, "Codet5+: Open code large language models for code understanding and generation," arXiv preprint arXiv:2305.07922, 2023.

[76] P. Manakul and M. J. F. Gales, "Sparsity and sentence structure in encoder-decoder attention of summarization systems," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.739 pp. 9359–9368.

[77] P. Manakul and M. J. F. Gales, "Long-span summarization via local attention and content selection," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.470 pp. 6026–6041.

[78] Y. Liu, J. Zhang, Y. Wan, C. Xia, L. He, and P. S. Yu, "HETFORMER: heterogeneous transformer with sparse attention for long-text extractive summarization," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.13 pp. 146–154.

[79] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A discourse-aware attention model for abstractive summarization of long documents," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), M. A. Walker, H. Ji, and A. Stent, Eds. Association for Computational Linguistics, 2018. [Online]. Available: https://doi.org/10.18653/v1/n18-2097 pp. 615–621.

[80] W. Xiao and G. Carenini, "Extractive summarization of long documents by combining global and local context," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/D19-1298 pp. 3009–3019.

[81] B. Pang, E. Nijkamp, W. Kryscinski, S. Savarese, Y. Zhou, and C. Xiong, "Long document summarization with top-down and bottom-up inference," CoRR, vol. abs/2203.07586, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2203.07586

[82] S. Cao and L. Wang, "HIBRIDS: attention with hierarchical biases for structure-aware long document summarization," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.58 pp. 786–807.

[83] Y. Dong, A. Romascanu, and J. C. K. Cheung, "Discourse-aware unsupervised summarization for long scientific documents," in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.eacl-main.93 pp. 1089–1102.

[84] Y. Liu, A. Ni, L. Nan, B. Deb, C. Zhu, A. H. Awadallah, and D. R. Radev, "Leveraging locality in abstractive text summarization," CoRR, vol. abs/2205.12476, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2205.12476

[85] Y. Zhang, A. Ni, Z. Mao, C. H. Wu, C. Zhu, B. Deb, A. H. Awadallah, D. R. Radev, and R. Zhang, "Summ$ˆn$: A multi-stage summarization framework for long input dialogues and documents," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.112 pp. 1592–1604.

[86] N. Gu, E. Ash, and R. H. R. Hahnloser, "Memsum: Extractive summarization of long documents using multi-step episodic markov decision processes," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.450 pp. 6507–6522.

[87] J. Ju, M. Liu, H. Y. Koh, Y. Jin, L. Du, and S. Pan, "Leveraging information bottleneck for scientific document summarization," in Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-emnlp.345 pp. 4091–4098.

[88] Z. Mao, C. H. Wu, A. Ni, Y. Zhang, R. Zhang, T. Yu, B. Deb, C. Zhu, A. H. Awadallah, and D. R. Radev, "DYLE: dynamic latent extraction for abstractive long-input summarization," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.118 pp. 1687–1698.

[89] X. Du, A. M. Rush, and C. Cardie, "GRIT: generative role-filler transformers for document-level event entity extraction," in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.eacl-main.52/ pp. 634–644.

[90] X. Du and C. Cardie, "Document-level event role filler extraction using multi-granularity contextualized encoding," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.714 pp. 8010–8020.

[91] H. Yang, D. Sui, Y. Chen, K. Liu, J. Zhao, and T. Wang, "Document-level event extraction via parallel prediction networks," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.492 pp. 6298–6308.

[92] R. Xu, T. Liu, L. Li, and B. Chang, "Document-level event extraction via heterogeneous graph-based interaction model with a tracker," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.274 pp. 3533–3546.

[93] X. Liu, Z. Luo, and H. Huang, "Jointly multiple events extraction via attention-based graph information aggregation," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, 2018. [Online]. Available: https://doi.org/10.18653/v1/d18-1156 pp. 1247–1256.

[94] B. Yang and T. M. Mitchell, "Joint extraction of events and entities within a document context," in NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, K. Knight, A. Nenkova, and O. Rambow, Eds. The Association for Computational Linguistics, 2016. [Online]. Available: https://doi.org/10.18653/v1/n16-1033 pp. 289–299.

[95] S. Li, H. Ji, and J. Han, "Document-level event argument extraction by conditional generation," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-main.69 pp. 894–908.

[96] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html pp. 5998–6008.

[97] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," in Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, A. Korhonen, D. R. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/p19-1580 pp. 5797–5808.

[98] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" 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, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019. [Online]. Available: http://papers.nips.cc/paper/9551-are-sixteen-heads-really-better-than-one pp. 14 014–14 024.

[99] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling BERT for natural language understanding," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.findings-emnlp.372 pp. 4163–4174.

[100] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from BERT into simple neural networks," CoRR, vol. abs/1903.12136, 2019. [Online]. Available: http://arxiv.org/abs/1903.12136

[101] W. Liu, P. Zhou, Z. Wang, Z. Zhao, H. Deng, and Q. Ju, "Fastbert: a self-distilling BERT with adaptive inference time," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.537 pp. 6035–6044.

[102] J. Qiu, H. Ma, O. Levy, W. Yih, S. Wang, and J. Tang, "Blockwise self-attention for long document understanding," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.findings-emnlp.232 pp. 2555–2565.

[103] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," CoRR, vol. abs/1904.10509, 2019. [Online]. Available: http://arxiv.org/abs/1904.10509

[104] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html

[105] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," CoRR, vol. abs/2012.07436, 2020. [Online]. Available: https://arxiv.org/abs/2012.07436

[106] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=rkgNKkHtvB

[107] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, "Efficient content-based sparse attention with routing transformers," Trans. Assoc. Comput. Linguistics, vol. 9, pp. 53–68, 2021. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/2405

[108] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," CoRR, vol. abs/2006.04768, 2020. [Online]. Available: https://arxiv.org/abs/2006.04768

[109] W. B. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," Contemporary mathematics, vol. 26, no. 189-206, p. 1, 1984.

[110] S. Har-Peled, P. Indyk, and R. Motwani, "Approximate nearest neighbor: Towards removing the curse of dimensionality," Theory Comput., vol. 8, no. 1, pp. 321–350, 2012. [Online]. Available: https://doi.org/10.4086/toc.2012.v008a014

[111] Y. Tay, D. Bahri, D. Metzler, D. Juan, Z. Zhao, and C. Zheng, "Synthesizer: Rethinking self-attention in transformer models," CoRR, vol. abs/2005.00743, 2020. [Online]. Available: https://arxiv.org/abs/2005.00743

[112] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh, "Nyströmformer: A nyström-based algorithm for approximating self-attention," CoRR, vol. abs/2102.03902, 2021. [Online]. Available: https://arxiv.org/abs/2102.03902

[113] C. K. I. Williams and M. W. Seeger, "Using the nyström method to speed up kernel machines," in Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper/2000/hash/19de10adbaa1b2ee13f77f679fa1483a-Abstract.html pp. 682–688.

[114] P. Drineas and M. W. Mahoney, "On the nyström method for approximating a gram matrix for improved kernel-based learning," J. Mach. Learn. Res., vol. 6, pp. 2153–2175, 2005. [Online]. Available: http://jmlr.org/papers/v6/drineas05a.html

[115] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," CoRR, vol. abs/2009.06732, 2020. [Online]. Available: https://arxiv.org/abs/2009.06732

[116] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler, "Long range arena: A benchmark for efficient transformers," CoRR, vol. abs/2011.04006, 2020. [Online]. Available: https://arxiv.org/abs/2011.04006

[117] A. Wang, R. Y. Pang, A. Chen, J. Phang, and S. R. Bowman, "Squality: Building a long-document summarization dataset the hard way," CoRR, vol. abs/2205.11465, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2205.11465

[118] U. Shaham, E. Segal, M. Ivgi, A. Efrat, O. Yoran, A. Haviv, A. Gupta, W. Xiong, M. Geva, J. Berant, and O. Levy, "SCROLLS: standardized comparison over long language sequences," CoRR, vol. abs/2201.03533, 2022. [Online]. Available: https://arxiv.org/abs/2201.03533

[119] Z. Deng, H. Peng, C. Xia, J. Li, L. He, and P. S. Yu, "Hierarchical bi-directional self-attention networks for paper review rating recommendation," in Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, D. Scott, N. Bel, and C. Zong, Eds. International Committee on Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.coling-main.555 pp. 6302–6314.

[120] K. Friedl, G. Rizos, L. Stappen, M. Hasan, L. Specia, T. Hain, and B. W. Schuller, "Uncertainty aware review hallucination for science article classification," in Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, ser. Findings of ACL, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., vol. ACL/IJCNLP 2021. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-acl.443 pp. 5004–5009.

[121] P. Li, Z. Wang, Z. Ren, L. Bing, and W. Lam, "Neural rating regression with abstractive tips generation for recommendation," in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, and R. W. White, Eds. ACM, 2017. [Online]. Available: https://doi.org/10.1145/3077136.3080822 pp. 345–354.

[122] H. P. Chan, W. Chen, and I. King, "A unified dual-view model for review summarization and sentiment classification with inconsistency loss," in Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds. ACM, 2020. [Online]. Available: https://doi.org/10.1145/3397271.3401039 pp. 1191–1200.

[123] L. Cheng, L. Bing, Q. Yu, W. Lu, and L. Si, "APE: argument pair extraction from peer review and rebuttal via multi-task learning," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.569 pp. 7000–7011.

[124] L. Wang and W. Ling, "Neural network-based abstract generation for opinions and arguments," in NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, K. Knight, A. Nenkova, and O. Rambow, Eds. The Association for Computational Linguistics, 2016. [Online]. Available: https://doi.org/10.18653/v1/n16-1007 pp. 47–57.

[125] S. Angelidis, R. K. Amplayo, Y. Suhara, X. Wang, and M. Lapata, "Extractive opinion summarization in quantized transformer spaces," Transactions of the Association for Computational Linguistics, vol. 9, pp. 277–293, 2021. [Online]. Available: https://aclanthology.org/2021.tacl-1.17

[126] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, "Dialogsum: A real-life scenario dialogue summarization dataset," in Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, ser. Findings of ACL, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., vol. ACL/IJCNLP 2021. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-acl.449 pp. 5062–5074.

[127] W. Xiao and G. Carenini, "Systematically exploring redundancy reduction in summarizing long documents," in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020, K. Wong, K. Knight, and H. Wu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.aacl-main.51/ pp. 516–528.

[128] C. Zhu, Y. Liu, J. Mei, and M. Zeng, "Mediasum: A large-scale media interview dataset for dialogue summarization," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-main.474 pp. 5927–5934.

[129] OpenAI, "Chatgpt-3.5-turbo," https://openai.com/research/, 2021, accessed: 2023-05-20.

[130] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Welleck, B. P. Majumder, S. Gupta, A. Yazdanbakhsh, and P. Clark, "Self-refine: Iterative refinement with self-feedback," 2023.

[131] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, July 2004. [Online]. Available: https://aclanthology.org/W04-1013 pp. 74–81.

[132] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," CoRR, vol. abs/1910.12840, 2019. [Online]. Available: http://arxiv.org/abs/1910.12840

[133] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "Summac: Re-visiting nli-based models for inconsistency detection in summarization," CoRR, vol. abs/2111.09525, 2021. [Online]. Available: https://arxiv.org/abs/2111.09525

[134] W. Zhao, M. Strube, and S. Eger, "Discoscore: Evaluating text generation with BERT and discourse coherence," CoRR, vol. abs/2201.11176, 2022. [Online]. Available: https://arxiv.org/abs/2201.11176

[135] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/n19-1423 pp. 4171–4186.

[136] J. Fu, S. Ng, Z. Jiang, and P. Liu, "Gptscore: Evaluate as you desire," CoRR, vol. abs/2302.04166, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.04166

[137] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: NLG evaluation using GPT-4 with better human alignment," CoRR, vol. abs/2303.16634, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.16634

[138] M. Gao, J. Ruan, R. Sun, X. Yin, S. Yang, and X. Wan, "Human-like summarization evaluation with chatgpt," 2023.

[139] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," J. Artif. Intell. Res., vol. 22, pp. 457–479, 2004. [Online]. Available: https://doi.org/10.1613/jair.1523

[140] N. Gu, E. Ash, and R. Hahnloser, "MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, May 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.450 pp. 6507–6522.

[141] T. Goyal, J. J. Li, and G. Durrett, "News summarization and evaluation in the era of GPT-3," CoRR, vol. abs/2209.12356, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2209.12356

[142] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[143] Y. Deng, A. Bakhtin, M. Ott, A. Szlam, and M. Ranzato, "Residual energy-based models for text generation," in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=B1l4SgHKDH

[144] T. He, B. McCann, C. Xiong, and E. Hosseini-Asl, "Joint energy-based model training for better calibrated natural language understanding models," in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.eacl-main.151 pp. 1754–1761.

[145] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," Predicting structured data, vol. 1, no. 0, 2006.

[146] S. Bhattacharyya, A. Rooshenas, S. Naskar, S. Sun, M. Iyyer, and A. McCallum, "Energy-based reranking: Improving neural machine translation using energy-based models," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.349 pp. 4528–4537.

[147] Q. Li, Y. Zhang, B. Li, L. Cao, and P. C. Woodland, "Residual energy-based models for end-to-end speech recognition," in Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-690 pp. 4069–4073.

[148] L. Qin, S. Welleck, D. Khashabi, and Y. Choi, "COLD decoding: Energy-based constrained text generation with langevin dynamics," CoRR, vol. abs/2202.11705, 2022. [Online]. Available: https://arxiv.org/abs/2202.11705

[149] G. Liu, Z. Yang, T. Tao, X. Liang, J. Bao, Z. Li, X. He, S. Cui, and Z. Hu, "Don't take it literally: An edit-invariant sequence loss for text generation," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.naacl-main.150 pp. 2055–2078.

[150] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," CoRR, vol. abs/2010.03759, 2020. [Online]. Available: https://arxiv.org/abs/2010.03759

[151] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. J. Crandall, and D. Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," CoRR, vol. abs/1610.02424, 2016. [Online]. Available: http://arxiv.org/abs/1610.02424

[152] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, 2018. [Online]. Available: https://doi.org/10.18653/v1/d18-1206 pp. 1797–1807.

[153] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," in Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Berlin, Germany: Association for Computational Linguistics, Aug. 2016. [Online]. Available: https://aclanthology.org/K16-1028 pp. 280–290.

[154] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "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, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.703 pp. 7871–7880.

[155] T. H. Nguyen, K. Cho, and R. Grishman, "Joint event extraction via recurrent neural networks," in NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, K. Knight, A. Nenkova, and O. Rambow, Eds. The Association for Computational Linguistics, 2016. [Online]. Available: https://doi.org/10.18653/v1/n16-1034 pp. 300–309.

[156] D. Wadden, U. Wennberg, Y. Luan, and H. Hajishirzi, "Entity, relation, and event extraction with contextualized span representations," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/D19-1585 pp. 5783–5788.

[157] Y. Lin, H. Ji, F. Huang, and L. Wu, "A joint neural model for information extraction with global features," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.713 pp. 7999–8009.

[158] W. A. Gale, K. W. Church, and D. Yarowsky, "One sense per discourse," in Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992. Morgan Kaufmann, 1992. [Online]. Available: https://aclanthology.org/H92-1045/

[159] F. Nan, R. Nallapati, Z. Wang, C. Nogueira dos Santos, H. Zhu, D. Zhang, K. McKeown, and B. Xiang, "Entity-level factual consistency of abstractive text summarization," in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, Apr. 2021. [Online]. Available: https://aclanthology.org/2021.eacl-main.235 pp. 2727–2733.

[160] O. Honovich, L. Choshen, R. Aharoni, E. Neeman, I. Szpektor, and O. Abend, "Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering," CoRR, vol. abs/2104.08202, 2021. [Online]. Available: https://arxiv.org/abs/2104.08202

[161] M. T. Ribeiro, C. Guestrin, and S. Singh, "Are red roses red? evaluating consistency of question-answering models," in Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, A. Korhonen, D. R. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, 2019. [Online]. Available: https://doi.org/10.18653/v1/p19-1621 pp. 6174–6184.

[162] P. Shi and J. Lin, "Simple BERT models for relation extraction and semantic role labeling," CoRR, vol. abs/1904.05255, 2019. [Online]. Available: http://arxiv.org/abs/1904.05255

[163] Q. Li, H. Ji, and L. Huang, "Joint event extraction via structured prediction with global features," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers. The Association for Computer Linguistics, 2013. [Online]. Available: https://aclanthology.org/P13-1008/ pp. 73–82.

[164] R. Huang and E. Riloff, "Modeling textual cohesion for event extraction," in Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada, J. Hoffmann and B. Selman, Eds. AAAI Press, 2012. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5113

[165] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," CoRR, vol. abs/1910.03771, 2019. [Online]. Available: http://arxiv.org/abs/1910.03771

[166] P. Ang, B. Dhingra, and L. W. Wills, "Characterizing the efficiency vs. accuracy trade-off for long-context NLP models," CoRR, vol. abs/2204.07288, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.07288

[167] L. Liu, X. Liu, J. Gao, W. Chen, and J. Han, "Understanding the difficulty of training transformers," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.463 pp. 5747–5763.

[168] N. Nangia and S. R. Bowman, "Listops: A diagnostic dataset for latent tree learning," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 2-4, 2018, Student Research Workshop, S. R. Cordeiro, S. Oraby, U. Pavalanathan, and K. Rim, Eds. Association for Computational Linguistics, 2018. [Online]. Available: https://doi.org/10.18653/v1/n18-4013 pp. 92–99.

[169] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. The Association for Computer Linguistics, 2011. [Online]. Available: https://www.aclweb.org/anthology/P11-1015/ pp. 142–150.

[170] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, "The ACL anthology network corpus," Lang. Resour. Evaluation, vol. 47, pp. 919–944, 2013. [Online]. Available: https://doi.org/10.1007/s10579-012-9211-2

[171] D. Linsley, J. Kim, V. Veerabadran, C. Windolf, and T. Serre, "Learning long-range spatial dependencies with horizontal gated recurrent units," in Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/ec8956637a99787bd197eacd77acce5e-Abstract.html pp. 152–164.

[172] A. Krizhevsky, G. Hinton et al., "Learning multiple layers of features from tiny images," 2009.

[173] D. P. Woodruff, "Sketching as a tool for numerical linear algebra," Found. Trends Theor. Comput. Sci., vol. 10, no. 1-2, pp. 1–157, 2014. [Online]. Available: https://doi.org/10.1561/0400000060

[174] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," CoRR, vol. abs/1011.3027, 2010. [Online]. Available: http://arxiv.org/abs/1011.3027

[175] N. Halko, P. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," SIAM Rev., vol. 53, no. 2, pp. 217–288, 2011. [Online]. Available: https://doi.org/10.1137/090771806

[176] N. Ailon and B. Chazelle, "Approximate nearest neighbors and the fast johnson-lindenstrauss transform," in Proceedings of the 38th Annual ACM Symposium on Theory of Computing, Seattle, WA, USA, May 21-23, 2006, J. M. Kleinberg, Ed. ACM, 2006. [Online]. Available: https://doi.org/10.1145/1132516.1132597 pp. 557–563.

[177] Y. Lu, P. S. Dhillon, D. P. Foster, and L. H. Ungar, "Faster ridge regression via the subsampled randomized hadamard transform," in Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/621bf66ddb7c962aa0d22ac97d69b793-Abstract.html pp. 369–377.

[178] Y. Yang, M. Pilanci, M. J. Wainwright et al., "Randomized sketches for kernels: Fast and optimal nonparametric regression," The Annals of Statistics, vol. 45, no. 3, pp. 991–1023, 2017.

[179] M. B. Cohen, J. Nelson, and D. P. Woodruff, "Optimal approximate matrix product in terms of stable rank," in 43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy, ser. LIPIcs, I. Chatzigiannakis, M. Mitzenmacher, Y. Rabani, and D. Sangiorgi, Eds., vol. 55. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016. [Online]. Available: https://doi.org/10.4230/LIPIcs.ICALP.2016.11 pp. 11:1–11:14.

[180] P. Li, T. Hastie, and K. W. Church, "Very sparse random projections," in Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006, T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, Eds. ACM, 2006. [Online]. Available: https://doi.org/10.1145/1150402.1150436 pp. 287–296.

[181] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast monte carlo algorithms for matrices I: approximating matrix multiplication," SIAM J. Comput., vol. 36, no. 1, pp. 132–157, 2006. [Online]. Available: https://doi.org/10.1137/S0097539704442684

[182] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[183] F. He, T. Liu, and D. Tao, "Control batch size and learning rate to generalize well: Theoretical and empirical evidence," 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, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/dc6a70712a252123c40d2adba6a11d84-Abstract.html pp. 1141–1150.

[184] J. Ding, S. Ma, L. Dong, X. Zhang, S. Huang, W. Wang, N. Zheng, and F. Wei, "Longnet: Scaling transformers to 1,000,000,000 tokens," 2023.

[185] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou, "Large language models can be easily distracted by irrelevant context," CoRR, vol. abs/2302.00093, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.00093

[186] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "Halueval: A large-scale hallucination evaluation benchmark for large language models," 2023.