Unlocking Structure Measuring: Introducing PDD, an Automatic Metric for Positional Discourse Coherence

Anonymous ACL submission

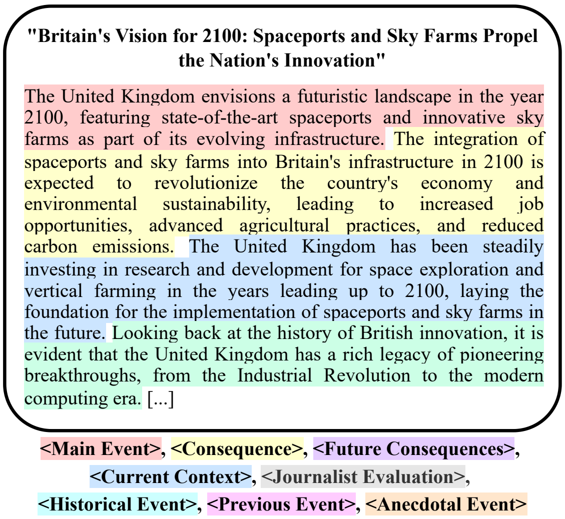
Abstract

Recent large language models (LLMs) have shown remarkable performance in aligning generated text with user intentions across various tasks. When it comes to long-form text generation, there has been a growing interest in generation from a discourse coherence perspective. However, existing lexical or semantic metrics such as BLEU, ROUGE, BertScore cannot effectively capture the discourse coherence. The development of discourse-specific automatic evaluation methods for assessing the output of LLMs warrants greater focus and exploration. In this paper, we present a novel automatic metric designed to quantify the discourse divergence between two long-form articles. Extensive experiments on three datasets from representative domains demonstrate that our metric aligns more closely with human preferences and GPT-4 coherence evaluation, outperforming existing evaluation methods.[1]

1 Introduction

Real-life texts often exhibit underlying structures. News articles, for instance, adhere to a specific narrative order, as illustrated in Fig. 1, employed by journalists to efficiently convey messages and improve reader experience. Despite recent advances in generation of fluent text, the generation of structurally coherent text remains an under-explored area. Follow the theory of functional discourse structure, elaborated in Appendix A.1, we leverage the discourse structure to model the coherence of long-form texts. Several recent works (Spangher et al., 2022; Liu et al., 2022) have addressed the problems of generating long-form text while following specific in-domain discourse schema.

While established automatic metrics such as BLEU (Papineni et al., 2002), ROUGE-L (Lin and Hovy, 2002) and BertScore (Zhang et al., 2019) exist for Natural Language Generation evaluation, they predominantly measure lexical n-gram overlaps or semantic similarities. The evaluation of structural coherence has been a long-existing challenge (Guan et al., 2021; Cho et al., 2019; Zhu and Bhat, 2020). A common baseline metric for measuring functional discourse structure is the exact match, which compares structure elements one-to-one at each exact position. However, this metric is notably sensitive to local variations and differences in the lengths of articles.



"Britain's Vision for 2100: Spaceports and Sky Farms Propel the Nation's Innovation"

The United Kingdom envisions a futuristic landscape in the year 2100, featuring state-of-the-art spaceports and innovative sky farms as part of its evolving infrastructure. The integration of spaceports and sky farms into Britain's infrastructure in 2100 is expected to revolutionize the country's economy and environmental sustainability, leading to increased job opportunities, advanced agricultural practices, and reduced carbon emissions. The United Kingdom has been steadily investing in research and development for space exploration and vertical farming in the years leading up to 2100, laying the foundation for the implementation of spaceports and sky farms in the future. Looking back at the history of British innovation, it is evident that the United Kingdom has a rich legacy of pioneering breakthroughs, from the Industrial Revolution to the modern computing era. [...]

<Main Event>, **<Consequence>**, **<Future Consequences>**, **<Current Context>**, **<Journalist Evaluation>**, **<Historical Event>**, **<Previous Event>**, **<Anecdotal Event>**

Figure 1: A news article example with discourse role annotations. The discourse schema follows the News discourse theory by Van Dijk (2013).

To address this gap, we propose a novel automatic model-free metric, Positional Discourse Divergence (PDD), specifically designed to evaluate the underlying discourse structure of articles in comparison to references. PDD partitions the sentences of an article into multiple position bins and calculates the divergence in discourse structures within each bin. This approach renders PDD resilient to various challenges encountered in long-form text generation, such as accommodating local variations and handling misaligned numbers of sentences.

[1] Our code will be available on GitHub.

To validate the effectiveness and generalizability of the PDD, we evaluate the inter-agreement with human evaluations and GPT-4 coherence evaluations on three representative datasets with different discourse schema: News Discourse (Choubey et al., 2020), Long-Form Question Answering (Xu et al., 2022a) and Recipe1M+ (Liu et al., 2022). Across all three domains, PDD demonstrates the highest agreement with human judgements on coherence.

2 Positional Discourse Divergence

Texts within a specific genre often exhibit similar patterns in their discourse sequences, albeit with some variations at a local level. In other words, the distribution of discourse roles is inherently tied to their approximate positions within the articles. For instance, News reports commonly present main events and their consequences at the beginning to capture the reader's interest, even though the precise order can differ. Likewise, recipes tend to follow a predictable structure, where the preparation of ingredients is generally mentioned first, followed by cooking actions towards the middle or end of the text.

Despite the fluency achieved by (large) Language Models, they struggle to organize discourse structures like humans. In Fig. 2, we observe disparities between the discourse distributions of model predictions and human-written references when the News articles are divided into 5 positional bins. To quantitatively capture these gaps, we introduce the Positional Discourse Divergence (PDD), denoted as D_{pos}, as an automatic metric. Equation 1 outlines the calculation for applying PDD to compare a *predicted* article against its corresponding *reference*:

$$D_{pos} = \frac{1}{N} \sum_{n=1}^{N} D_{KL}(p^n(r) + \boldsymbol{\epsilon} || q^n(r) + \boldsymbol{\epsilon}) \quad (1)$$

Firstly, both articles are segmented into N positional bins. Note the number of bin, N, should be smaller than the number of sentences in both the reference and the prediction. We denote $p^n(r)$ to represent the distribution of discourse role r for the *reference* in the n-th position bin, and $q^n(r)$ to represent the distribution for the *generated* article. These discourse distributions are calculated by the frequency density of the discourse roles within each bin. These discourse distributions are calculated by the frequency density of the discourse roles within each bin. Then, for each bin n, the KL divergence between the discourse distributions is calculated. [2] To address the sparsity in the discourse distribution of a single article, small-value terms, denoted as $\boldsymbol{\epsilon}$, are introduced. This addition helps in avoiding instances of zero probabilities in the distribution.

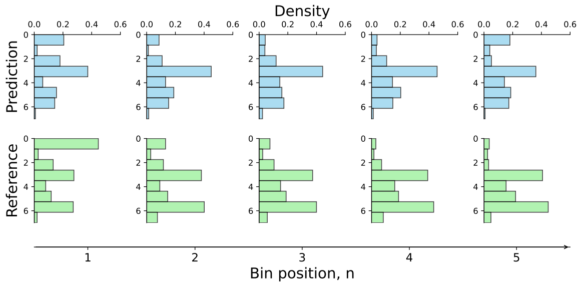


Figure 2: Positional discourse distribution comparisons (N=5). Top row: The discourse distribution of model predictions on News Discourse test set (Llama2-7b, finetuned on Kaggle All the News). Bottom row: Test set reference distributions.

To compute PDD or other discourse measurements like exact match, it is inevitable to employ a discourse role classifier for labeling both prediction and reference articles. An off-the-shelf discourse classifier, trained on human-annotated data with a defined schema (e.g., the News Discourse dataset for news domain), can serve this purpose. Further information regarding the discourse classifiers is provided in Appendix B.

2.1 Interpreting the Metric

2.1.1 Set vs. Individual Predictions

Much like the BLEU score, the Positional Discourse Divergence can be applied to a set of text, including both the set of model predictions and the set of reference articles. The underlying assumption is that all articles within a given set adhere to similar discourse structures, for example, being News articles of the same sports genre. Consequently, the discourse distributions of this set of articles offer a more accurate estimate of the discourse distribution specific to that genre.

In the assessment of a single predicted article against a reference set, the focus is on how well the article aligns with the target genre. In contrast, when comparing a prediction set against a reference set, the evaluation exams the model's overall ability to generate content of that specific genre.

[2] Due to the asymmetry nature of KL divergence, $D_{KL}(P||Q)$ is interpreted as the information divergence of Q against P. Accordingly, we employ $q^n(r)$ to denote the discourse distributions of the predictions and $p^n(r)$ for the reference.

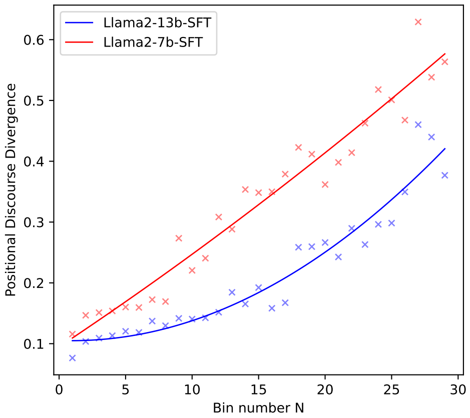


Figure 3: Positional Discourse Divergence vs. Bin number (N) for predictions by two language models on the News Discourse test set. Training details in Appendix C. Curves represent best-fit quadratic curves.

2.1.2 Bin Number

The bin number, N, plays a crucial role in determining the sensitivity of PDD to local variations, as illustrated in Fig. 3. Therefore, the behavior of PDD varies with the choice of N. Intuitively, a larger N implies lower tolerance for local perturbations. When N equals the number of sentences, the PDD is essentially equivalent to the exact match metric. Whereas, when N equals 1, it describes the overall discourse role distribution gaps between the prediction to the reference.

To illustrate, we fine-tuned Llama2-7b and Llama2-13b (Touvron et al., 2023) and compared their predictions with the reference News articles. The details of the supervised fine-tuning process are explained in Appendix C. The PDD curves, illustrating the performance with different choices of bin numbers, are presented in Fig. 3. The gaps in performance are effectively captured by the disparities between the two PDD curves.

3 Metric Validation

To validate the efficacy of the Positional Discourse Divergence metric, we evaluate its agreement with human assessments, and GPT-4 on article coherence. Additionally, we compare PDD against baseline automatic metrics, such as exact match, BLEU, and BertScore. To assess generalizability, we conduct this validation across three different domains, each characterised by distinct human annotated, sentence-level discourse schemas:

(I) News. We utilize the News Discourse dataset (Choubey et al., 2020), comprising 802 documents across four genres and three media sources. The average number of sentences per article is 14.6. Manual annotations for the News Discourse dataset follow the theory of functional discourse schema proposed by Van Dijk (1988, 2013). This schema defines discourse based on eight types of relations between each sentence and the main event.

(II) Long-form QA. Long-Form Question Answering (LFQA) involves providing comprehensive answers composed of multiple sentences. Xu et al. (2022a) proposed an discourse ontology of six sentence-level functional roles also following the theory of functional discourse structure. The discourse annotations are collected on three recent LFQA datasets (ELI5 (Fan et al., 2019), WebGPT (Nakano et al., 2022), and Natural Questions (Kwiatkowski et al., 2019)). A total of 640 answer paragraphs were released, with an average of 6.1 sentences per paragraph.

(III) Recipes. We adopt the discourse schema proposed by Liu et al. (2022) which includes seven discourse roles based on cooking actions specifically designed for recipes. They annotated the Recipe1M+ dataset (Moryossef et al., 2019; Marín et al., 2021) with a rule-based annotation system following the proposed schema. The Recipe1M+ contains over 1M textual recipes and ingredients.

For further information regarding dataset details and schema definitions, please refer to Appendix E.

3.1 Comparison with Other Metrics

We validate the effectiveness of our metric, PDD, by assessing its inter-agreement with human evaluations and GPT-4 coherence evaluations. The human evaluation setup details can be found in Appendix D. In our comparison, PDD is evaluated alongside several automatic metrics, including exact match, BLEU (Papineni et al., 2002), ROUGE-L (Lin and Hovy, 2002) and BertScore (Zhang et al., 2019). Notably, only PDD and exact match focus on directly measuring discourse structure, while the others are designed for assessing n-gram or semantic similarity.

As rating long-form articles with absolute scores is a relatively complicated and subjective task, we instead ask evaluators compare two perturbed variations of the original reference article. Cohen's Kappa is computed between the metrics and evaluators based on these preference annotations. We create two variations in the way that prevents resulting PDD values from exhibiting a heavy left-tail issue and ensuring a more accurate kappa estimation. In Variation 1, we randomly shuffle all the

Metrics	Human			GPT-4		
	News Discourse	LFQA	Recipe1M+	News Discourse	LFQA	Recipe1M+
Exact Match	0.26	0.29	0.43	0.42	0.24	0.25
ROUGE-L	0.30	0.24	0.39	0.44	0.19	0.26
BLEU	0.24	0.31	0.46	0.48	0.28	0.32
BertScore	**0.45**	0.42	0.63	0.63	**0.42**	0.68
PDD	0.42	**0.49**	**0.66**	**0.65**	0.38	**0.71**

Table 1: Cohen's Kappa with human and GPT-4 coherence evaluations. For News Discourse, bin number $N = 8$ was used, while for LFQA and Recipe1M+, $N = 3$. Human evaluation involves 50 randomly selected example pairs for each dataset, while GPT-4 evaluation uses 300 pairs.

sentences, whereas in Variation 2, we initially segment the article into a randomly selected number of bins and then shuffle sentences only within their respective positional bins.

In Tab. 1, we report the Kappa with both human and GPT-4 coherence evaluations. The details of the prompt and survey templates are shown in Appendix G. Our PDD metric demonstrates consistent good agreement (0.4-0.6) in News Discourse and LFQA, achieving substantial agreement (>0.7) on Recipe1M+ dataset. The notable performance on the Recipe1M+ dataset can be attributed to the strong order-dependent nature of recipes: A shuffled question-answer format may be challenging to understand, but a disordered recipe is nearly incomprehensible.

Another observation on News Discourse dataset, indicates Kappas with human evaluations are generally lower than those with GPT-4. This discrepancy is likely due to the much longer length of news articles compared to question answering and recipe datasets, posing a more challenging task for human readers.

Our PDD metric significantly outperforms baseline metrics of Exact Match, Rouge-L and BLEU, while achieving comparable Kappa with the BertScore. We attribute the good performance of BertScore to its ability in carrying textual knowledge from the pre-trained BERT. In our experiment setup, both Variation 1 and 2 are shuffled from the same articles. Consequently, the metrics based on n-gram and semantic similarities can effectively distinguish examples closer to the original version and therefore achieve high kappa values. However, when comparing the discourse structure between two different articles of the same genre, they are likely to have very different n-gram or semantic content while maintaining a similar discourse structure. In this case, only Exact Match and PDD can capture the divergence between discourse structure.

3.2 Discussion

Our experimental findings yield the following noteworthy observations:

- The behavior of PDD, as indicated by the formula in Eq. 1, converges towards Exact Match as the chosen bin number increases. Conversely, with a smaller value of N, PDD consistently outperforms Exact Match in terms of kappa. This observation validates our initial hypothesis that permitting a certain level of local variation does not detrimentally impact the overall reader experience.
- PDD consistently exhibits high kappa scores across diverse domains, emphasizing the significance of preserving discourse structure in text across various subject areas.
- PDD is specifically designed to evaluate the underlying discourse structure. It is not only simple and model-free, eschewing reliance on pre-trained language models, but also interpretable because of its intrinsic use of KL divergence.

4 Conclusion

In conclusion, the exploration of text generation with natural underlying structure remains a significantly under-explored domain. Addressing this gap, we introduced PDD, a simple and model-free metric designed to assess discourse structure. By quantifying the divergence between discourse distributions within position bins, PDD exhibits robust agreement with human and GPT-4 coherence evaluations across three representative domains, outperforming a range of baseline metrics. Our hope is that PDD will stimulate future research endeavors focused on unraveling the intricacies of underlying structure in text generation.

Limitations

Discourse classifier requirement We note that our PDD requires a discourse classifier when applied to model predictions. Although this necessity is inevitable in evaluating the discourse structure alignment for any other metric such as Exact Match, it underscores the dependence on annotated data with the target discourse schema for training.

Choice of bin number N The choice of bin number will affect the performance of the PDD. However, the ideal choice of N may vary for different articles: The optimal number of sections the article should be segmented into. While trends may exist within specific genres or datasets, in general, it requires certain level of domain expertise to determine the optimal bin number.

References

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. 2019. Towards coherent and cohesive long-form text generation. In *Proceedings of the First Workshop on Narrative Understanding*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence. 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)*, pages 6379–6393, Online. Association for Computational Linguistics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. Can transformer models measure coherence in text: Re-thinking the shuffle test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing & Management*, 27(1):55–81.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51.

Yinhong Liu, Yixuan Su, Ehsan Shareghi, and Nigel Collier. 2022. Plug-and-play recipe generation with content planning. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 223–234, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Javier Marín, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2021. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6):468–487.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Alexander Spangher, Yao Ming, Xinyu Hua, and Nanyun Peng. 2022. Sequentially controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6848–6866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Teun A Van Dijk. 1988. News analysis. *Case Studies of International and National News in the Press. New Jersey: Lawrence*.

Teun A Van Dijk. 2013. *News as discourse*. Routledge.

Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022a. How do we answer complex questions: Discourse structure of long-form answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3556–3572, Dublin, Ireland. Association for Computational Linguistics.

Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022b. How do we answer complex questions: Discourse structure of long-form answers. *arXiv preprint arXiv:2203.11048*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

A Background and Related Works

A.1 Discourse Structure

Discourse structure investigates the organization of language into larger units like paragraphs, sections, and complete articles. In this work, we focus on the communicative functions within entire articles served by those linguistic units. Therefore, texts from different domains are characterized by different discourse schemas, as their linguistic units also play different functional roles. The discourse roles of scientific papers or experimental abstracts (Liddy, 1991; Mizuta et al., 2006) include background, methodology, experiments and findings. In the domain of long-form question answering Xu et al. (2022b), the discourse function of each sentence can be answer, summary, example and so on. Liu et al. (2022) developed a discourse schema for recipes based on actions and controlled the generation process according to the predicted discourse sequences. The explicit functional discourse structure of news reports was addressed (Van Dijk, 2013; Choubey et al., 2020) by defining roles based on their relations with the main event, such as consequence and journalist evaluation.

Multiple established frameworks also proposed different definition of discourse structure, which focus on how each linguistic unit relates to each other through discourse connectives, such as causal, temporal, etc. For instance, Rhetorical Structure Theory, RST (Mann and Thompson, 1988), seeks to identify rhetorical relations between text segments and form a hierarchical organization of discourse. The Penn Discourse Treebank, PDTB (Prasad et al., 2008), defines its schema based on low-level discourse connectives presented in the text.

A.2 NLG Metrics

Traditional NLG metrics, such as BLEU (Papineni et al., 2002) and ROUGE-L (Lin and Hovy, 2002), measure lexical n-gram overlaps to assess fluency, but they have limitations in capturing semantic similarity. Later works tried to improve the hard lexical matching with soft word embedding matching (Ng and Abrecht, 2015) or stemming and synonym matching (Lavie and Agarwal, 2007).

Recently, by leveraging contextual embeddings from BERT (Kenton and Toutanova, 2019), a series of metrics can successfully capture semantic similarity with references or even textual quality without references(Zhao et al., 2019; Zhang et al., 2019; Yuan et al., 2021). However, as BERT is argued that can only capture limited discourse information (Koto et al., 2021; Laban et al., 2021; Beyer et al., 2021), they are not suitable for evaluating the discourse structure in long texts.

DiscoScore (Zhao et al., 2023) is a BERT-based metric, specifically designed to model local discourse coherence for summarization and document-level machine translation tasks. By leveraging Center theory (Grosz et al., 1995), they modelled discourse similarity by focus frequency and transitions.

B Discourse Classifier

A discourse classifier is usually a lightweight language model trained on sentence-discourse role pairs. Here we report the classifier performance achieved:

For the News domain, we train a DistilBERT (Sanh et al., 2019) as the discourse role classifier on the News Discourse training set and evaluated on the validation set using human-annotated gold labels. The classifier achieves an accuracy of 67%.

In the Recipe domain, the reported performance of the discourse role classifier, by Liu et al. (2022), achieves an accuracy of 92%. It was a a DistilBERT model trained on the Recipe1M+ training set and evaluated on the validation set using silver annotations generated by a rule-based system.

For LFQA, Xu et al. (2022a) achieved an accuracy of 54% by a T5-large model, which shows comparable performances to human. The classifier was trained and tested on the ELI5 dataset.

C LLM SFT Details

We fine-tuned two language models, the 8-bit LoRa versions of Llama2-7b and Llama2-13b, using Kaggle All the News train set comprising 42.4K samples after filtering. The models receive news headlines as input and aim to generate the corresponding news articles. Training employed consistent hyperparameters: a learning rate of 3×10^{-4}, 2 epochs, LoRa parameters $r = 8$, $\alpha = 16$, and dropout set at 0.05. The models were trained on a single RTX a6000 48GB, requiring 12 and 23 hours for Llama2-7b and Llama2-13b, respectively.

D Human Evaluation Details

The human evaluation was conducted using Amazon Mechanical Turk (MTurk). We obtained three preference annotations for each example pair from native English-speaking crowd workers. The final results were determined based on the majority preference among the three evaluations. Crowd workers were compensated at a rate of 15 pounds per hour for their participation in the evaluation process.

E Discourse Schema

The definition of the discourse schema we used for news articles:

- **Main Event**: The major subject of the news article.
- **Consequence**: An event or phenomenon that is caused by the main event.
- **Previous Event**: A specific event that occurred shortly before the main event.
- **Current Context**: The general context or world state immediately preceding the main event.
- **Historical Event**: An event occurring much earlier than the main event.
- **Future Consequences**: An analytical insight into future consequences or projections.
- **Journalist Evaluation**: A summary, opinion or comment made by the journalist.
- **Anecdotal Event**: An event that is uncertain and cannot be verified. The primary purpose is to provide more emotional resonance to the main event.

The definition of the discourse schema for LFQA:

- **Organizational sentence**: An organizational sentence is to inform the reader how the answer will be structured.
- **Answer summary**: An answer sentence that plays a summary role, which can often suffice by themselves as the answer to the question.
- **Answer**: Answer sentences which explain or elaborate on the summary.
- **Example**: The example provided in answers, which discussed a particular entity or concept that is different from the rest of the answer sentences.
- **Auxiliary Information**: Provide information that are related, but not directly asked in the question.
- **Miscellaneous**: Various other roles that shows up in human answers, such as the limitation of the answer or the source of the answer.

The definition of the discourse schema we used for recipes:

- **Pre-processing** means the preparations of ingredients or cooker.
- **Mixing** includes actions of combining one or more ingredients together.
- **Transferring** is for the actions of moving or transferring food or intermediate food to a specific place.
- **Cooking** represents the actual cooking actions, which could vary drastically across different recipes.
- **Post-processing** usually refers to the following up actions after the ‘cooking’ stage, such as ‘cooling down’, ‘garnish’.
- **Final** refers to the last few actions before serving the food or the serving action itself.
- **General** includes the rest of actions which cannot be classified into the above categories.

F Data Preprocessing

For News Discourse, we filtered the dataset based on the following conditions:

- Containing special characters: @, [, +.
- Having total number of words over 800 or below 100.
- Containing random comments.
- Containing more than two reports.

Then we pre-process the data by

- Removing extra space.
- Removing reporting source.
- Removing journalist names.
- Removing emoji.

For Recipe1M+, we filter it based on the following codintions:

- Containing irrelevant information, such as advertisements, reviews and comments.
- Having total number of words over 300 or below 50.
- Duplicate recipes.

For LFQA, we filtered out all model generated answer paragraphs, because they contain sentences that do not have assigned discourse roles.

G Evaluation Templates

Human evaluation instruction Below, we provide the instruction example for human evaluation on the News Discourse dataset, where evaluators were directed to express their preference. The instructions for LFQA and Recipe1M+ are similar, with certain domain-specific keywords substituted, such as News headline becoming Recipe title.

“ Read the two versions of the news for the given headline and rank their coherence following the guideline below.

Coherence guidelines:

1. Flow of Sentences: Evaluate how well the sentences transition from one to another. A fluent text should have seamless connections between sentences.

2. Logical Organization: Evaluate how well the sentences are organized and the ideas are conveyed. A coherent text should have a clear and precise structure.

General guidelines:

1. Be Objective: Please focus on the coherence of writing, not the content or opinions expressed.

2. Please rate which one is preferred between the two versions.

News headline: {headline}

Version 1: {version1}

Version 2: {version2} ”

GPT-4 evaluation prompt Below, we provide the prompt template for GPT-4 coherence evaluation on the News Discourse dataset. Although the GPT-4 was instructed to rate with scores, but the scores are converted to preference later. The instructions for LFQA and Recipe1M+ are similar, with certain domain-specific keywords substituted, such as News headline becoming Recipe title.

“ Pretend you are a human reader. Please evaluate the coherence of the two given news articles. Guideline:

1. Rate on a scale of 1 to 10, where 1 represents very low coherence, and 10 indicates very high coherence.

2. Consider the flow of ideas and the ordering of sentences. A highly coherent article should have a better sentence ordering.

3. Must return ratings in JSON format only: {"score1": [your rating for version 1], "score2": [your rating for version 2]}

News headline: [headline]

News version 1: [version1]

News version 2: [version2]

Rating: ”