Semantically Enriched Text Generation for QA through Dense Paraphrasing

Timothy Obiso, Bingyang Ye, Kyeongmin Rim, and **James Pustejovsky**
Department of Computer Science
Brandeis University
Waltham, Massachusetts
{timothyobiso, byye, krim, jamesp}@brandeis.edu

Abstract

Large Language Models (LLMs) are very effective at extractive language tasks such as Question Answering (QA). While LLMs can improve their performance on these tasks through increases in model size (via massive pretraining) and/or iterative on-the-job training (one-shot, few-shot, chain-of-thought), we explore what other less resource-intensive and more efficient types of data augmentation can be applied to obtain similar boosts in performance. We define multiple forms of Dense Paraphrasing (DP) and obtain DP-enriched versions of different contexts. We demonstrate that performing QA using these semantically enriched contexts leads to increased performance on models of various sizes and across task domains, without needing to increase model size.

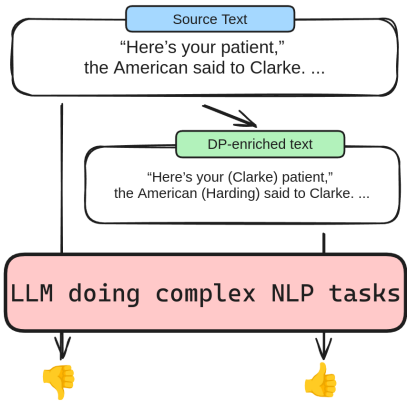

Figure 1: Dense Paraphrasing and LLM. We hypothesize the economy of natural language plays an important role in the degraded performance of LLMs on Natural Language Processing (NLP) tasks, and by augmenting the context text at prompt time by DP, we can boost performance.

1 Introduction

In this paper, we explore different methods of semantically enriching reference texts to improve the performance of Large Language Models (LLMs) on downstream tasks, particularly Question Answering (QA). There are a number of common ways to increase the performance of LLMs on these tasks: fine-tuning, few-shot prompting, and data augmentations. Traditionally, data augmentation is done to increase the amount of training data available with the hope that more data will lead to better performance.

In the context of LLM usage, we propose data augmentation in terms of enriching the context text in a prompt. To do this, we augment the data used as a reference for the QA task to be more semantically informative; this is Dense Paraphrasing (DP). Then, we use the new text as the reference and perform the task. We see noticeable improvements in automatic and human metrics on the answers obtained by models using DP-enriched text.

Our contributions are as follows:

- We formalize multiple forms of Dense Paraphrasing: Anaphora and Coreference Dense Paraphrasing and Semantic Role Labelling Dense Paraphrasing and propose computationally efficient ways of obtaining these paraphrases, avoiding multiple LLM calls.
- DP-enriched text outperforms the original text on automatic metrics and human evaluation.
- Dense Paraphrasing improves performance on models of all sizes: this includes Llama3 8B and Llama3 70B.

We use smaller models such as the spaCy coreference model[1] and the Verb Net parser (Gung, 2020; Gung and Palmer, 2021) to generate DP-enriched text. We then perform the QA task using the original text and the DP-enriched text and compare our results. This pipeline is based on the illustration in Figure 1. We have made all of our code publicly available on a public code repository.[2]

[1] en_coreference_web_trf
[2] https://github.com/brandeis-llc/dpqa

2 Related Work

Many transformer-based models have proven themselves well-suited to QA tasks. The best models have traditionally involved BERT or RoBERTa (Ju et al., 2019; Wu et al., 2019). Other approaches have involved ensembling the responses of multiple models (Ju et al., 2019; Zhu et al., 2018). The newest approaches use LLMs such as GPT (Brown et al., 2020) to perform these tasks.

To improve LLM performance at these tasks under zero-shot, one-shot, few-shot, and fine-tuning conditions, a number of data augmentation strategies have been proposed, summarized by Chen et al. (2023). These methods include EDA (Easy Data Augmentation) (Wei and Zou, 2019), SeemSeek (Kim et al., 2022), AMR-DA (Shou et al., 2022), Back-translation (Sennrich et al., 2016), Dialog Inpainting (Dai et al., 2022), and AutoConv (Li et al., 2023).

The examples generated from these DA steps are generally used to train models of smaller size (<1B). In this work, we use the augmented examples in the zero-shot prompt to perform the QA task.

This approach is based on query reformulation techniques, widely used in the field of information retrieval (Bruza and Dennis, 1997) and database management systems (Rajaraman et al., 1995), and then adopted for more complex NLP tasks under different names such as "Decontextualization" (Choi et al., 2021) or "Dense Paraphrasing" (Tu et al., 2022, 2023).

3 Types of Dense Paraphrasing

Tu et al. (2023) define Dense Paraphrasing as generating text that "reduces ambiguity while also making explicit the underlying semantics that [are] not expressed in the economy of sentence structure".

In this work, we define two forms of DP. These methods all saturate the text with additional information yet differ in what information is added. The following Dense Paraphrasing methods are ways to clarify various semantic relations in a text.

3.1 Anaphora and Coreference

One way to perform DP is by clarifying which entity is being referred to whenever an anaphoric or coreferential expression is used. We refer to this process as Anaphora and Coreference Dense Paraphrasing (A/C DP). This process duplicates names next to all entity expressions, reinserting information available from prior context.

(1) S1: *"Here's your patient," the American said to Clarke. "We expect you to cure him, and you had better get to work at once."*

S2: *"Here's your **(Clarke)** patient **(Blake)**," the American **(Harding)** said to Clarke **(Clarke)**. "We expect you **(Clarke)** to cure him **(Blake)**, and you **(Clarke)** had better get to work at once."*

A/C DP is obtained by adding in the name of the entity being referred to after each referential expression. We link together entity chains grouped by a coreference model through the spaCy package. The first mention of an entity is the name that is duplicated next to each mention of the entity.

3.2 Semantic Role Labelling

We explore another method of DP which focuses on event participant roles. By employing Semantic Role Labeling (SRL) to recover the predicate-argument structure of the sentence, we hope the model can better understand the wh-questions: "who did what to whom", "when", and "where" (Màrquez et al., 2008). Specifically, we use an off-the-shelf SRL tool VerbNet Parser (Gung, 2020; Gung and Palmer, 2021) for this purpose. Compared to traditional SRL systems, the VerbNet Parser infuses knowledge from the English Lexical resource VerbNet (Brown et al., 2019, 2022) for enhanced disambiguation of the predicate. Further, its thematic roles are more semantically informative than those in traditional SRL.

For our task, we run the VerbNet Parser on all contexts and questions. We extract the syntax roles and insert them back into the sentence immediately following the text span they correspond to. In example 2, S1 is the original sentence and S2 is the DP-enriched sentence. The VerbNet Parser detects that the predicate in S1 is "sit" and the matching frame requires a Theme and a Location. Then it extracts the value from S1 for these two thematic roles.

(2) S1: *My grandfather was sitting in the backyard.*
S2: *My grandfather **(Theme)** was sitting **(Verb)** in the backyard **(Location)**.*

3.3 Combining DPs

A text can also be passed through multiple layers of DP. The result is a text that contains multiple

	No DP		A/C DP		SRL DP	
	EM	F1	EM	F1	EM	F1
Llama3 8B	43.3	57.0	42.5	56.6	**49.9**	**63.6**
Llama3 70B	45.9	61.9	45.7	61.3	**47.3**	**64.9**

Table 1: The impact of Dense Paraphrasing on CoQA

types of semantic information presented alongside the original text. We experiment with combining the results of A/C DP and Semantic Role Labelling Dense Paraphrasing (SRL DP) into a dually-DP-enriched text. We present these results in the Appendix.

4 Experiments

We conduct experiments on the Conversational Question Answering (CoQA) dataset (Reddy et al., 2019).

4.1 Data

Conversational Question Answering (CoQA) is a prominent dataset designed for the task of conversational QA. The task is designed to examine the models' capability to understand the dialogue flow and respond to a sequence of questions based on a given passage. CoQA contains 127k questions with answers, obtained from 8k conversations about text passages from seven diverse domains, including News, Literature, Exams, etc.

CoQA is designed to model conversational QA and was created in an interactive mode where a questioner asks a sequence of questions based on a passage while a responder answers them. This design is an example of multi-turn dialogue for datasets. This setup ensures that the questions asked are more natural than conventional QA pairs. By training models on these datasets, we hope to foster the development of models that can handle dialogue flow and maintain context across longer amounts of text.

4.2 Methods

We first obtain DP-enriched versions of the context paragraphs of each example from CoQA dataset. We use the publicly available Llama3 8B and Llama3 70B models (AI@Meta, 2024). We run both of our Llama3 8B and 70B experiments on NVIDIA RTX A6000 with 48GB vRAM (300W power supplied).

We perform the QA task without any DP as a baseline. Figure 2 shows the prompt we feed to LLMs. We repeat the task using the same prompt but with contexts enriched with A/C DP and SRL DP.

Context: Once upon a time, in a barn near a farm house, there lived a white little kitten named Cotton ...

Q_1: What color was Cotton?
A_1: white

....

Q_{k-1}: What did the other cats do when Cotton emerged from the bucket of water??
A_{k-1}: licked her face

Question: Did they want Cotton to change the color of her fur?

Answer succinctly with ONLY words or phrases EXACTLY from the text

Figure 2: An example prompt of Question K for the QA task

4.3 Results

Our results are summarized in Table 1. A/C DP slightly hurts general performance while SRL DP greatly improves it. Using SRL DP-enriched text caused an increase of 6.6 percentage points each in EM (exact match) and F1 with Llama3 8B. We report an increase of 1.4 and 3.0 percentage points for EM and F1 using SRL DP-enriched text with Llama3 70B.

The CoQA development set contains five domains: children's stories, literature, middle-high school stories, news, and Wikipedia. Table 2 shows how while A/C DP can help in some domains, its improvements are not consistent enough throughout the entire dataset and hurt performance in many cases.

As shown in Table 2, SRL DP improves every metric for Llama3 8B and all but one metric for Llama3 70B. This indicates that DP, specifically SRL DP, can induce better performance at extractive language tasks across domains using models of various sizes. This motivates further use of SRL DP as a data augmentation step to increase the performance of LLMs.

We also performed a round of human evaluation on the first thirty stories in the CoQA development set. These results are shown in Table 3. From this, we can see that both types of DP led to improved performance.

4.4 Error Analysis

We classify the common errors made by our models into the following categories:

Reasoning Error The models usually make this type of error when the answer cannot be directly

		CS		Lit		MHS		News		Wiki		All	
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Llama3 8B	No DP	43.8	58.8	34.6	48.3	42.5	56.7	46.2	58.5	49.6	63.0	43.3	57.0
	A/C DP	39.7	56.4	35.8	49.5	38.0	52.8	46.8	60.4	52.0	63.8	42.5	56.6
	SRL DP	**48.2**	**63.8**	**41.9**	**54.5**	**47.8**	**62.8**	**52.7**	**65.4**	**58.7**	**71.5**	**49.9**	**63.6**
Llama3 70B	No DP	43.3	60.8	**42.6**	**57.8**	43.2	60.4	46.3	62.0	54.0	68.4	45.9	61.9
	A/C DP	42.5	60.5	41.9	57.2	40.8	57.3	46.1	61.5	**57.0**	69.9	45.7	61.3
	SRL DP	**45.0**	**65.1**	40.3	**57.8**	**45.4**	**63.4**	**48.4**	**65.3**	**57.0**	**73.0**	**47.3**	**64.9**

Table 2: CoQA by Domain

	No DP	A/C DP	SRL DP
Llama3 8B	64.5	66.5	**70.5**
Llama3 70B	76.1	77.7	**79.7**

Table 3: Human Evaluation on the first 30 QA sets (251 questions) from CoQA reported as accuracy

extracted through keywords, or it requires some extra reasoning to understand the question. For example, if the story describes how a duck is sad about herself being different from the rest of her family, these models struggle to answer a question that asks, "Is the duck happy about it?"

Intention Error This happens when the model fails to understand the intention of a yes-no question and, instead, answers with extractive information. Although the information may be relevant, the answer to a yes-no should be "Yes" or "No". For example, the question is "did they write back?" and the model answers "write a note to her." The answer can be understood from this text but it is not an answer to the question. A prompt that includes specific instructions for yes-no questions may alleviate this error.

Follow-up Error When answering a follow-up question, the models may not be able to detect that this is a continuation of the question asked previously. For instance, consider the question "What were they like" which refers to the man's clothes according to the preceding question. The model answers "tough-as-nail", which describes the man's character in movies. The wrong interpretation of "they" in the question causes the answer to be true but not relevant. This error can be greatly alleviated through A/C DP where coreference resolution would replace a pronoun with the actual entity and recover the previous context.

In addition to these errors, both automatic evaluation, EM and F1, and human evaluation, accuracy, will miss some semantically correct answers. These include the yes-no errors as well as answers that are accurate but have taken the wording of the story and rephrased it while maintaining the same meaning.

5 Discussion

In natural language, even some required arguments of event predicates can be omitted due to the economy of sentence structure. This can pose a challenge for downstream tasks like QA. VerbNet Parser can not only extract existing thematic roles of a sentence but can also indicate whether a thematic role is missing. Given that information, we could perform saturation of missing roles by recovering the covert arguments for each event and place these arguments back into the text. This is another form of DP, Frame Saturation Dense Paraphrasing (FS DP), similar to what is done manually in Rim et al. (2023).

Future work can explore other types of DP that provide semantic information in plain text or even other means of doing so. All of our experiments were conducted under zero-shot conditions. Our results motivate experiments and research using DP with few-shot prompting and fine-tuning. In particular, we recommend fine-tuning an LLM on large amounts of DP-enriched training data before performing downstream tasks on DP-enriched text.

We also note that SRL DP improves performance throughout the conversational exchange. As shown in Figures 3 and 4, SRL DP boosts both metrics. However, as prompt length increases, the DP-enriched text seems the same performance drop as the original text. This semantic enrichment improves the ability of models of multiple sizes to answer questions and draw conclusions where the necessary information is spread across a very long text, up to 25 questions and answers long.

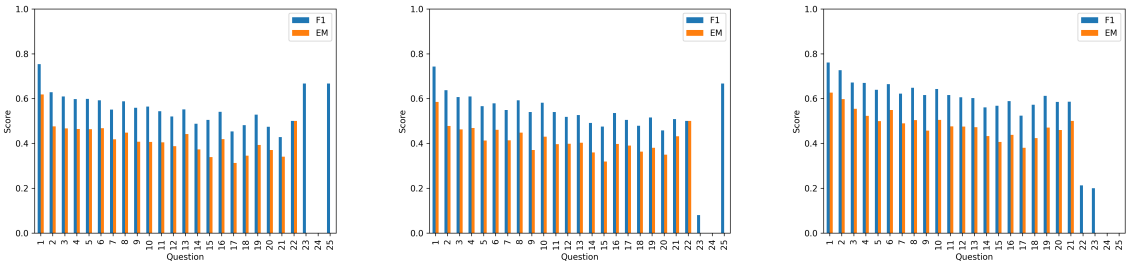

Figure 3: F1 and EM for the 8B model for No DP, A/C DP, and SRL DP

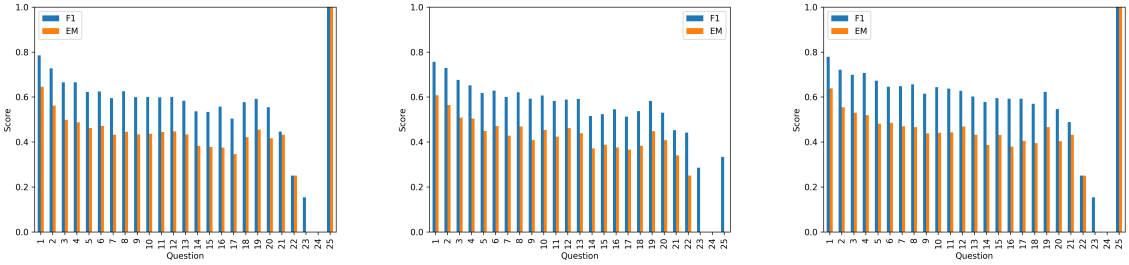

Figure 4: F1 and EM for the 70B model for No DP, A/C DP, and SRL DP

6 Conclusion

From our experiments, we conclude that Dense Paraphrasing, specifically Anaphora and Coreference Dense Paraphrasing (A/C DP) and Semantic Role Labelling Dense Paraphrasing (SRL DP), can help LLMs perform extractive tasks such as QA. A/C DP and SRL DP both enrich texts with semantic information that language models can use to more accurately perform downstream extractive tasks.

Limitations

Our experiments and evaluation were only limited to the CoQA dataset. These texts are all of a similar, finite length. The dataset only covers a limited number of domains and is only in English.

Ethics Statement

Any risks related to the unsupervised use of LLMs are present here. We do not perform manual or automatic checks or filters on the data we have evaluated or in our system. While there are safeguards in place in LLMs to protect from offensive content and bias (Liang et al., 2021; Roy et al., 2023; Sahoo et al., 2024), they are not perfect (Wang et al., 2024). During our limited human evaluation, we did not come across any biased, harmful, or offensive content in the dataset or generated by our system.

Acknowledgements

This work was supported in part by NSF grant 2326985 to James Pustejovsky. The opinions and views reported herein are those of the authors alone.

References

AI@Meta. 2024. Llama 3 model card.

Susan Windisch Brown, Julia Bonn, James Gung, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2019. VerbNet representations: Subevent semantics for transfer verbs. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 154–163, Florence, Italy. Association for Computational Linguistics.

Susan Windisch Brown, Julia Bonn, Ghazaleh Kazeminejad, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2022. Semantic representations for nlp using verbnet and the generative lexicon. *Frontiers in artificial intelligence*, 5:821697.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Peter Bruza and Simon J. Dennis. 1997. Query reformulation on the internet: Empirical data and the hyperindex search engine. In *RIAO Conference*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data

augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International conference on machine learning*, pages 4558–4586. PMLR.

James Gung. 2020. *Abstraction, Sense Distinctions and Syntax in Neural Semantic Role Labeling*. University of Colorado at Boulder.

James Gung and Martha Palmer. 2021. Predicate representations and polysemy in VerbNet semantic parsing. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 51–62, Groningen, The Netherlands (online). Association for Computational Linguistics.

Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on conversational question answering. *arXiv preprint arXiv:1909.10772*.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Towards more realistic generation of information-seeking conversations.

Siheng Li, Cheng Yang, Yichun Yin, Xinyu Zhu, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. 2023. AutoConv: Automatically generating information-seeking conversations with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1751–1762, Toronto, Canada. Association for Computational Linguistics.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Lluís Màrquez, Xavier Carreras, Kenneth C Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue.

Anand Rajaraman, Yehoshua Sagiv, and Jeffrey D. Ullman. 1995. Answering queries using templates with binding patterns (extended abstract). In *Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS '95, page 105–112, New York, NY, USA. Association for Computing Machinery.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Kyeongmin Rim, Jingxuan Tu, Bingyang Ye, Marc Verhagen, Eben Holderness, and James Pustejovsky. 2023. The coreference under transformation labeling dataset: Entity tracking in procedural texts using event models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12448–12460, Toronto, Canada. Association for Computational Linguistics.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Nihar Ranjan Sahoo, Ashita Saxena, Kishan Maharaj, Arif A. Ahmad, Abhijit Mishra, and Pushpak Bhattacharyya. 2024. Addressing bias and hallucination in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 73–79, Torino, Italia. ELRA and ICCL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ziyi Shou, Yuxin Jiang, and Fangzhen Lin. 2022. AMR-DA: Data augmentation by Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3082–3098, Dublin, Ireland. Association for Computational Linguistics.

Jingxuan Tu, Kyeongmin Rim, Eben Holderness, Bingyang Ye, and James Pustejovsky. 2023. Dense paraphrasing for textual enrichment. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 39–49, Nancy, France. Association for Computational Linguistics.

Jingxuan Tu, Kyeongmin Rim, and James Pustejovsky. 2022. Competence-based question generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1521–1533, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024. Do-not-answer: Evaluating safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911, St. Julian's, Malta. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Jindou Wu, Yunlun Yang, Chao Deng, Hongyi Tang, Bingning Wang, Haoze Sun, Ting Yao, and Qi Zhang. 2019. Sogou Machine Reading Comprehension Toolkit. *arXiv e-prints*, page arXiv:1903.11848.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

A Appendix

	No DP		A/C + SRL DP		SRL + A/C DP	
	EM	F1	EM	F1	EM	F1
Llama3 8B	43.3	57.0	40.8	55.6	40.1	52.4
Llama3 70B	45.9	61.9	45.1	60.9	39.1	54.6

Table 4: CoQA results of text enriched with multiple forms of DP

We also ran the QA task using text that has been enriched by both SRL DP and A/C DP. Including both types of information in the same format hurt performance on the CoQA dataset regardless of the order they were added in. These results are shown in Table 4