Generating Interpretations of Policy Announcements

Andreas Marfurt[1,2,*], **Ashley Thornton**[3], **David Sylvan**[3], and **James Henderson**[2]

[1]Lucerne University of Applied Sciences and Arts, Switzerland
[2]Idiap Research Institute, Switzerland
[3]Graduate Institute of International and Development Studies, Switzerland
[*]Correspondence to: andreas.marfurt@hslu.ch

Abstract

Recent advances in language modeling have focused on (potentially multiple-choice) question answering, open-ended generation, or math and coding problems. We look at a more nuanced task: the interpretation of statements of political actors. To this end, we present a dataset of policy announcements and corresponding annotated interpretations, on the topic of US foreign policy relations with Russia in the years 1993 up to 2016. We analyze the performance of finetuning standard sequence-to-sequence models of varying sizes on predicting the annotated interpretations and compare them to few-shot prompted large language models. We find that 1) model size is not the main factor for success on this task, 2) finetuning smaller models provides both quantitatively and qualitatively superior results to in-context learning with large language models, but 3) large language models pick up the annotation format and approximate the category distribution with just a few in-context examples.

1 Introduction

State-of-the-art language models are evaluated on multiple-choice question answering (e.g. MMLU; Hendrycks et al., 2021), math problems (e.g. GSM8k; Cobbe et al., 2021), or coding (e.g. HumanEval; Chen et al., 2021). These benchmarks do not provide much insight for the use and analysis of such models in the humanities and social sciences.

In this paper, we present a dataset on an important issue in the humanities and social sciences, namely interpretation. In this case, our concern is with how newspaper articles characterize policy announcements (press releases, Q&A sessions, interviews, etc.). These interpretations are carefully annotated by political scientists to give them structure (through labeling spans of text with a category such as *act* or *motive*) and to provide additional background knowledge as comments. We then train language models on the task of generating the annotated interpretations when shown the announcement.

```
[STD SENTENCE START] On the eve of [ACTOR START]
President Bush's (USA) [ACTOR END] [ACT START]
arrival here [REFERENCE START] to sign
[REFERENCE END] a nuclear arms reduction treaty
(The US and Russia will sign START II, a new
arms control agreement) [ACT END] ,
[RUSSIA LINK START] President Boris N. Yeltsin
is being criticized for pushing through an
accord that some say serves American interests
and confirms Russia's subordinate status in a
post-Communist world (The US and Russia will
sign START II, a new arms control agreement)
[RUSSIA LINK END] . [STD SENTENCE END]
```

Figure 1: Example annotated interpretation with a highlighted act, accompanied by a comment explaining necessary background knowledge.

We compare sequence-to-sequence models with large language models (LLMs), and find that model size is not indicative of task performance. We achieve better results by finetuning the comparatively much smaller sequence-to-sequence models than by few-shot prompting LLMs. Our code, data and models are available on GitHub[1].

2 Related Work

Language models have previously been used to interpret figurative language (Liu et al., 2022; Chakrabarty et al., 2022), contracts (Hoffman and Arbel, 2023; Wang, 2024), and building regulations (Fuchs et al., 2023). We provide a novel dataset on interpreting policy announcements.

Although using large language models to perform interpretation seems to have become more popular recently, the analysis of policy statements has focused mainly on either monetary policy (Doh et al., 2021; Lee et al., 2021; Marfurt et al., 2022)

[1]https://github.com/idiap/policy-interpretations

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 513–520
November 16, 2024.

or legislative speech (Goplerud, 2021). Regarding the latter, a dataset for sentiment analysis of political debates, ParlVote (Abercrombie and Batista-Navarro, 2020) has been created. Other work has focused on analyzing speech acts in political debates (Reinig et al., 2024). The policy announcements in this paper differ from political debates by being performed on behalf of a single actor and encompassing a wide range of issue areas; the interpretations of those announcements often presume background knowledge on the part of readers, which is challenging to capture.

Finally, using annotations to generate semi-structured outputs with language models has been used in Galactica (Taylor et al., 2022) to annotate paper citations and specific character sequences (DNA, amino acids), and to interpret economic policy announcements by the Federal Reserve Bank (Marfurt et al., 2022). We deem the format of the latter useful for our task as well and will employ it in the following section.

3 Dataset

The dataset concerns the foreign policy relations of the United States of America with respect to Russia in the years from 1993 up to 2016. A team of political scientists has curated announcements (press releases, Q&A sessions, interviews, etc.) and corresponding interpretations (partial, complete, or multiple sentences of *New York Times* articles). Details of the dataset's creation can be found in Appendix A. For annotation, we follow the guidelines of Marfurt et al. (2022). We define the mandatory annotation categories of a *standardized sentence* to be *act*, *actor*, and *reference*, with the same meaning as in previous work, although in this case, the actor is almost always the United States. We add a mandatory annotation category *Russia link*, which marks the connection of the announcement to Russia. We import the optional categories (*attribution*, *evidence*, *motive*, *scope*) without any changes. Comments, which make explicit newspaper readers' presumed background knowledge, are added in parentheses after text annotated as *act* or *Russia link*.

We convert the annotated interpretations into text-only format by inserting start and end markers (Taylor et al., 2022; Marfurt et al., 2022). An example from the training set can be seen in Figure 1, and statistics for the dataset are listed in Table 1.

	Train	Valid	Test
Source announcements	2116	250	264
Target interpretations	3360	404	378
Target std sentences	5240	636	579
Mean source words	6923	6967	6979
Mean target words	223	223	220

Table 1: Dataset statistics.

4 Experiments

In our experiments, we compare different approaches to solve our proposed task. To generate the target interpretations, we compare finetuning sequence-to-sequence models with in-context learning with large language models. We select T5 (Raffel et al., 2020), Flan-T5 (Chung et al., 2022), BART (Lewis et al., 2020) as our sequence-to-sequence models for their track record on previous text generation tasks. As large language models, we choose LLaMA-3.1 (8B) (Dubey et al., 2024) and Gemma-2 (9B) (Team et al., 2024). We decided to only use local language models for reasons of reproducibility.

4.1 Metrics

We aim to capture a diverse signal from the model interpretations. We measure the lexical similarity of generations to reference interpretations with ROUGE (Lin, 2004). We do this both on the complete generations including the annotation markers (termed *ROUGE-full*) and just on the generated words (*ROUGE-words*). We measure semantic similarity with BERTScore (Zhang et al., 2020), and use baseline rescaling[2]. For lexical diversity and to avoid repetitions, we analyze distinct bigrams as the number of unique bigrams divided by the total number of generated bigrams. To evaluate how often models copy from the source document, we compute the fraction of novel bigrams in the generated text compared to the source document. Finally, we aim for a more detailed evaluation of the similarity of predicted acts to ground truth acts. To this end, we measure ROUGE-2 (high correlation with human judgments for summaries (Fabbri et al., 2021)) for the contents of the annotated acts.

4.2 Training Details

Training is only performed for sequence-to-sequence models. They are finetuned for 20 epochs

[2]Evaluation hash: `roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.44.0)-rescaled`

Model	Parameters	ROUGE-full			ROUGE-words			BERTScore	Distinct	Novel	ROUGE-2
		R-1	R-2	R-L	R-1	R-2	R-L	(rescaled)	2-grams	2-grams	(acts)
References									58.95%	82.51%	
Finetuned seq2seq models											
T5 (base)	222M	40.42	13.86	27.89	35.90	13.01	24.73	19.24	50.01%	62.01%	3.77
Flan-T5 (base)	247M	40.79	13.79	28.00	36.30	12.93	24.93	19.04	50.98%	58.47%	7.74
BART (large)	406M	43.87	16.28	31.01	38.48	15.33	27.64	22.18	50.22%	70.35%	10.49
T5 (large)	737M	40.62	13.87	27.85	36.01	12.95	24.79	20.76	51.49%	69.06%	4.00
Flan-T5 (large)	783M	40.36	14.44	28.88	35.27	13.46	25.20	19.12	49.60%	63.04%	7.91
5-shot prompted LLMs											
LLaMA-3.1	8.03B	22.36	3.06	13.21	21.98	3.35	13.64	-18.01	46.31%	76.37%	0.35
LLaMA-3.1 (instruct)	8.03B	33.08	7.10	21.41	29.46	7.44	20.22	12.01	48.27%	73.08%	2.95
Gemma-2	9.24B	21.74	3.46	13.42	21.52	3.69	13.78	-17.80	38.97%	72.32%	0.41
Gemma-2 (instruct)	9.24B	31.40	6.14	20.63	26.37	6.62	18.79	7.99	57.20%	72.34%	2.12

Table 2: Test set results.

with early stopping (we try stopping both based on the validation loss or validation ROUGE score). For each of the models, we performed hyperparameter optimization on the learning rate. We started 10 training runs per model with varying learning rates (1e-3 to 1e-6). We trained each model on a single RTX A6000 GPU with an effective batch size of 8. We use the Adam optimizer (Kingma and Ba, 2015) and warm up the learning rate for 2 epochs. As our models can only process inputs of 1024 tokens, we filter the announcements with an oracle that selects the sentences that maximize the ROUGE-2 score when compared to the annotation (Liu and Lapata, 2019). Because of the lack of available pretrained long-context models, we leave ingesting the entire announcement into the model for future work.

4.3 Inference Details

When generating with the sequence-to-sequence models, we use beam search with 5 beams. We generate at least 100 and at most 512 tokens. We use n-gram blocking with $n = 6$ (Paulus et al., 2018).

For LLMs, we provide 5 in-context examples of an announcement with a corresponding interpretation, taken from the training set. We then prompt with the evaluation announcement. The announcements are prefixed with *Announcement:*, and the interpretations with *Interpretation:*. We generate with nucleus (top_p) sampling (Holtzman et al., 2020). We vary the temperature (0.5, 0.7, 1.0, 2.0) and the top_p (0.5, 0.8, 0.9, 0.95). We compare loading the model's weights in float32 with bfloat16 (16-bit precision shows a small performance drop). For instruction-tuned LLaMA, we also try using a system prompt[3] (not available for Gemma). In total, we try 12 hyperparameter combinations per LLM. The best settings can be found in Appendix B.

5 Results

The results in Table 2 show that the BART (large) model performs best among the models we tried. It scores the highest on lexical similarity (ROUGE) and semantic similarity (BERTScore), and achieves reasonable diversity and novelty of generated text. Appendix C shows an example output. We now present our main findings from these results.

Increasing model size does not improve results. We experimented with different-sized sequence-to-sequence models. We cannot make out a general trend in the change of performance due to model size. The best-performing model BART is of medium size. Additionally, LLMs do not reach the performance of the finetuned smaller models on this task.

Instruction tuning helps in-context learning. For both LLaMA and Gemma, the instruction-tuned versions massively outperform the base models on all metrics that measure similarity with the reference interpretations. The negative rescaled scores of the base LLMs suggest that content-wise, the LLMs' generations are less semantically similar to the ground truth than two randomly drawn sentences from Common Crawl (cf. Zhang et al., 2020). It seems that instruction tuning is a necessary ingredient of LLM training to enable in-context learning on this task.

Instruction tuning drastically shortens outputs. While not listed in Table 2, we also find that instruction-tuned models generate shorter outputs (less than half the tokens than the base models, and even shorter than the sequence-to-sequence

[3] "You are a chatbot that analyzes political announcements and replies with a coded interpretation of its main points."

Model	Std sent	Act	Motive	Evidence	Russia link	Correct format
References	1.55 (± 0.85)	1.57 (± 0.89)	0.21 (± 0.57)	0.60 (± 0.93)	1.58 (± 0.92)	99.41%
T5 (base)	1.54 (± 0.66)	0.62 (± 0.53)	0.00 (± 0.00)	0.00 (± 0.00)	0.15 (± 0.37)	85.67%
Flan-T5 (base)	1.48 (± 0.59)	1.11 (± 0.67)	0.00 (± 0.00)	0.05 (± 0.25)	0.37 (± 0.56)	94.56%
BART (large)	1.78 (± 0.59)	1.54 (± 0.60)	0.06 (± 0.26)	0.56 (± 0.89)	1.29 (± 0.60)	96.68%
T5 (large)	1.42 (± 0.54)	0.47 (± 0.52)	0.00 (± 0.00)	0.00 (± 0.00)	0.30 (± 0.50)	85.78%
Flan-T5 (large)	1.70 (± 0.61)	1.23 (± 0.64)	0.00 (± 0.06)	0.06 (± 0.25)	0.41 (± 0.58)	89.58%
In-context examples	1.2	1.2	0	0	1.2	
LLaMA-3.1	0.13 (± 0.33)	0.21 (± 0.47)	0.00 (± 0.00)	0.12 (± 0.39)	0.16 (± 0.39)	61.53%
LLaMA-3.1 (instruct)	1.03 (± 0.70)	0.70 (± 0.62)	0.00 (± 0.00)	0.39 (± 0.69)	0.61 (± 0.60)	89.60%
Gemma-2	0.23 (± 0.47)	0.17 (± 0.39)	0.00 (± 0.00)	0.08 (± 0.30)	0.13 (± 0.33)	78.72%
Gemma-2 (instruct)	1.17 (± 0.68)	0.59 (± 0.51)	0.00 (± 0.00)	0.07 (± 0.27)	0.63 (± 0.51)	95.55%

Table 3: Selected annotation counts with standard deviation on the test set.

models). They also produce many more annotation start and end marker tokens (around 10% of total tokens), whereas base models generate only around 2%. References contain 12.4% of these special tokens.

Table 3 shows the counts and standard deviation for a selection of annotation categories. Again, the BART model matches the reference distribution the closest (except for standardized sentences). Motives are underrepresented in all model outputs. In the last column, we also report if models correctly open and close annotations with matching start and end markers[4].

Only BART follows the reference category distribution. All sequence-to-sequence models and the instruction-tuned LLMs generate more than one standardized sentence on average. However, except for BART, models seldom generate all the mandatory categories of an interpretation. Moreover, BART generates the correct format more often than any other model.

Instruction tuning is vital for learning the format and distribution. For both LLaMA and Gemma, there is a major difference between the base model and the instruction-tuned model. Instruction tuning both allows the models to pick up the distribution of annotation categories and the annotation format with start and end markers much better, reaching similar performance as the sequence-to-sequence models. For some categories (evidence, Russia link), they generate more annotations than the T5 and Flan-T5 models. All this is achieved with only 5 in-context examples.

[4]References are not 100% correct since if they are too long (we used 512 tokens), the matching end markers get cut off.

LLMs generate categories that are not in the in-context examples. Surprisingly, we found that LLMs also generated categories not present in our in-context examples. Especially evidence is generated quite frequently, particularly by LLaMA (instruct). A natural explanation is that the LLMs must have been pretrained on a similar dataset that contained those or similar annotations. If this is the case, it is still interesting to see that both LLMs transfer that pretraining knowledge so readily. The annotation category scope, which is not shown in Table 3, appears twice in the in-context examples, yet is generated fewer times by all LLMs except the base LLaMA. This, however, also means that even though our dataset has not been released yet, performance on it will depend on whether the used models have been pretrained on similar datasets.

On top of the annotation categories present in our dataset, LLaMA also generates the additional categories *location* and *source*, while Gemma generates *location* and *organization*. They are, however, very rare, appearing at most 4 times for our total of 264 announcements in the test set.

6 Conclusion

We presented a new dataset on generating interpretations for policy announcements concerning US foreign policy with respect to Russia for the years 1993 to 2016. We evaluated common language models on this generation task and found that fine-tuned sequence-to-sequence models, specifically BART, outperformed few-shot prompted large language models.

Interesting directions for future work are investigating long-context methods that can access the entire announcement when writing the interpretation, and a more structured approach to generating the individual categories and their contents. We also

hope to see whether models trained on this dataset can be successfully transferred to other tasks and domains.

Limitations

We discuss limitations of our work in the following.

Limitations of the dataset. While the target interpretations are carefully curated by human experts, the source announcements are extracted from PDFs and websites. Especially text extracted from the former may include artifacts, such as additional or missing whitespace and punctuation caused by unusual formatting, or only partially extracted text.

Limitations of evaluated methods. We believe that the performance of LLMs could be improved in various ways. Since the presented task is rather complex, more few-shot examples could be given as additional context for the model to adapt. In some cases, Gemma even asked for more information in its response, e.g. "I am still under development and learning to interpret complex text. Can you please provide me with more context or specify what you would like me to do with this text?" Furthermore, parameter-efficient finetuning of LLMs (e.g. low-rank adaptation; Hu et al., 2022) may further improve their results. We leave these investigations to future work.

Limitations of evaluation metrics. In this paper, we evaluate models with established automatic metrics for text generation. For the task of generating interpretations, metrics comparing a candidate to a reference interpretation by lexical or semantic similarity will naturally miss the more subtle aspects. An evaluation that extracts the main characteristics of the interpretation in a broader context is interesting for future work.

Acknowledgments

This work was supported as a part of the grant Automated interpretation of political and economic policy documents: Machine learning using semantic and syntactic information, funded by the Swiss National Science Foundation (grant number CRSII5_180320), and led by the co-PIs James Henderson, Jean-Louis Arcand and David Sylvan.

References

Gavin Abercrombie and Riza Batista-Navarro. 2020. ParlVote: A corpus for sentiment analysis of political debates. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France. European Language Resources Association.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Taeyoung Doh, Sungil Kim, and Shu-Kuei Yang. 2021. How you say it matters: Text analysis of fomc statements using natural language processing. *Economic Review-Federal Reserve Bank of Kansas City*, 106(1):25–40.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Stefan Fuchs, Michael Witbrock, Johannes Dimyadi, and Robert Amor. 2023. Using large language models for the interpretation of building regulations. In *13th Conference on Engineering, Project and Production Management*.

Max Goplerud. 2021. Methods for analyzing parliamentary debates. In *The Politics of Legislative Debates*, chapter 5. Oxford University Press.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

David A. Hoffman and Yonathan A. Arbel. 2023. Generative interpretation. *New York University Law Review*, 99.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. 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*.

Jean Lee, Hoyoul Luis Youn, Nicholas Stevens, Josiah Poon, and Soyeon Caren Han. 2021. Fednlp: An interpretable nlp system to decode federal reserve communications. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2560–2564, New York, NY, USA. Association for Computing Machinery.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Andreas Marfurt, Ashley Thornton, David Sylvan, Lonneke van der Plas, and James Henderson. 2022. A corpus and evaluation for predicting semi-structured human annotations. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 262–275, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *Preprint*, arXiv:2211.09085.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Brydon T Wang. 2024. Prompts and large language models: A new tool for drafting, reviewing and interpreting contracts? *Law, Technology and Humans*, 6(2):88–106.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Dataset Creation

We outline the details of dataset creation. First, the New York Times archives were searched for articles on the foreign policy of the US with respect to Russia. In these articles, domain experts searched

Model	Early stopping	Max LR model	Max LR LM head
T5 (base)	Loss	1e-4	1e-3
Flan-T5 (base)	ROUGE	1e-4	1e-4
BART (large)	ROUGE	1e-5	1e-4
T5 (large)	ROUGE	1e-4	1e-4
Flan-T5 (large)	ROUGE	1e-4	1e-3

Table 4: Best hyperparameter settings for sequence-to-sequence models based on the validation set.

Model	Temperature	Top_p
LLaMA-3.1	1.0	0.9
LLaMA-3.1 (instruct)	0.7	0.5
Gemma-2	0.7	0.95
Gemma-2 (instruct)	0.7	0.9

Table 5: Best hyperparameter settings for LLMs based on the validation set.

for partial, complete or multiple sentences that contain all the information for the required annotation categories described in Section 3. All categories were then marked and potentially commented on to surface the readers' necessary background knowledge. Then, they are validated by a senior domain expert.

B Hyperparameter Settings

We list the optimal hyperparameter settings for each of our models in Tables 4 and 5.

C Example Outputs

In Table 6, we show a source announcement and the corresponding reference interpretation and interpretations of BART and Gemma-2 (instruct). In this example, BART focuses on a different part of the speech than the reference interpretation, which could have nevertheless been picked up in a an article. Gemma gets the main point right but becomes too repetitive. BART uses annotation categories in the right places but does not close them correctly in the second sentence. Gemma closes all annotations correctly but places them on parts of the text that do not match.

Source announcement

From State Department Dispatch, Vol. 4, No. 3, 1993: Chemical Weapons Convention Signing Ceremony Secretary Eagleburger Remarks upon signing the Chemical Weapons Convention, Paris, France, January 13, 1993.

It is fitting that we meet to sign this historic Chemical Weapons Convention in a city where, 4 years ago, the international community appealed for the strengthening of norms against chemical warfare. I am pleased to be in Paris, and I am especially pleased to represent my President, George Bush, a man who, over the course of the past decade, launched some of the key initiatives which helped to make this agreement possible. He and all those responsible can take pride in an achievement whose revolutionary scope and impact we can recognize today without having to await the verdict of history. But such has been the amazing record of the past few years. We have seen the international community liberate itself from half a century of gridlock and paralysis and move beyond the rhetoric of democracy to achieve real democracy; move beyond the rhetoric of detente to achieve real peace; and move beyond the rhetoric of disarmament to achieve real reductions in weapons of mass destruction. The Chemical Weapons Convention we sign today does more than simply reduce a class of arms or mitigate against their proliferation. This convention mandates a worldwide non-discriminatory ban on an entire class of weapons of mass destruction–the only class of such weapons that has been widely used in combat. By the radical terms of this agreement, all signatory states forswear the possession, production, stockpiling, transfer, and, indeed, the use of chemical weapons; and all signatories must destroy all chemical weapons and chemical weapons production facilities in their possession. Moreover, the convention's strict verification regime, which accommodates legitimate commercial and sovereign interests, sets an innovative standard for future multilateral agreements. The international community is virtually united in support of the objectives of the Chemical Weapons Convention. However, there must be truly global adherence if the convention is to achieve its purpose and if doubts are to be eliminated over the commitment and intentions of those who fail to sign, ratify, and fully comply with its terms. Nowhere is this more important today than in the Middle East, a region which over the past 30 years has been home to more active chemical weapons programs–and which has seen more chemical weapons use–than any other part of the world. It is, therefore, particularly disappointing that so many Middle Eastern states are absent from this ceremony today. The fact of the matter is that linking this convention to other issues cannot affect the fate of those issues, but it will surely undermine the effect of this treaty in the one region most exposed to the danger of chemical weapons–namely, the Middle East. The point, I believe, is to tackle the challenge of weapons of mass destruction wherever we can, whenever we can. I would, therefore, urge the members of the Arab League to seize this opportunity and sign the Chemical Weapons Convention. Doing so would be a step toward, and not away from, making the Middle East a zone free of all weapons of mass destruction, as called for by President Mubarak of Egypt. Today's ceremony is only the beginning of the work which lies ahead. Next month, the Preparatory Commission will meet in The Hague [the Netherlands] to work out the important and detailed provisions for implementing the convention. The United States is fully committed to the success of those efforts, which will require the same broad support and participation which produced the successful convention itself. As I indicated at the beginning, the past few years have been a remarkably creative period of international achievement. Perhaps not coincidentally, I believe that President Bush's passage across the international scene has equally been one of tangible achievement, particularly in terms of the issue most important to the fate and future of the planet–the issue of weapons of mass destruction. George Bush's legacy will include landmark treaties–START [Strategic Arms Reduction Treaty] I, START II, and CFE [Conventional Armed Forces in Europe]–as well as diplomatic efforts which paid non-proliferation dividends in Africa, South America, the Middle East, and here in Paris today. But he knows, as all of us must know, that what we have accomplished to date will matter little unless we are prepared to confront the even greater proliferation dangers we most certainly will face in the years to come.

Reference interpretation

[STD SENTENCE START] [ACT START] [RUSSIA LINK START] [ACTOR START] The United States (USA) [ACTOR END] and Russia were among those signing [REFERENCE START] today (The US signed the Chemical Weapons Convention) [ACT END] (Russia also signed the Chemical Weapons Convention) [RUSSIA LINK END] [REFERENCE END] . Speaking on behalf of the United States, which has already begun destroying its chemical-weapons stockpiles, Secretary of State Lawrence S. Eagleburger noted that [EVIDENCE START] the treaty mandates a global ban on the only class of weapons of mass destruction "that has been widely used in combat." [EVIDENCE END] [STD SENTENCE END]

[STD SENTENCE START] In [EVIDENCE START] what has been hailed as the first attempt to eliminate an entire category of weapons of mass destruction [EVIDENCE END] , [ACT START] [RUSSIA LINK START] more than 120 nations, including [ACTOR START] the United States (USA) [ACTOR END] and Russia, [REFERENCE START] began today [REFERENCE END] to sign an agreement to ban the production, stockpiling and use of chemical weapons (The US signed the Chemical Weapons Convention) [ACT END] (Russia also signed the Chemical Weapons Convention) [RUSSIA LINK END] . [STD SENTENCE END]

Candidate interpretation by BART

[STD SENTENCE START] [ACTOR START] Secretary of State Warren Christopher (United States) [ACTOR END] [ACT START] [REFERENCE START] urged [REFERENCE END] the Arab League today to join the [RUSSIA LINK START] Chemical Weapons Convention (Russia is a signatory of the Chemical Weapons Convention) [RUSSIA LINK END], which bans chemical weapons (The US urged the Arab League to sign the Chemical Weapons Convention and thus become part of the WMD regime) [ACT END]. [STD SENTENCE END]

[STD SENTENCE START] [ACTOR START] Mr. Christopher (United States [ACTOR END] [REFERENCE START] said [REFERENCE END], [ACT START] " [EVIDENCE START] The point, I believe, is to tackle the challenge of weapons of mass destruction wherever we can, whenever we can [EVIDENCE END]. Nowhere is this more important today than in the Middle East, a region which over the past 30 years has been home to more active chemical weapons programs – and which has seen more chemical weapons use – than any other part of the world [EVIDENCE END]." [STD SENTENCE END]

ROUGE-full (1/2/L): 55.22/20.34/41.75, BERTScore (rescaled): 15.20, novel 2-grams: 46.43%

Candidate interpretation by Gemma-2 (instruct)

[STD SENTENCE START] [RUSSIA LINK START] Mr. Eagleburger (USA) [RUSSIA LINK END] stressed that [ACT START] [REFERENCE START] the United States [REFERENCE END] [ACTOR START] is committed to [ACTOR END] working with all nations to achieve the universal ratification and implementation of the Chemical Weapons Convention (The United States is committed to working with all nations to achieve the universal ratification and implementation of the Chemical Weapons Convention) [ACT END] (The United States is committed to working with all nations to achieve the universal ratification and implementation of the Chemical Weapons Convention) [STD SENTENCE END] .

ROUGE-full (1/2/L): 39.33/14.35/30.13, BERTScore (rescaled): 12.65, novel 2-grams: 74.19%

Table 6: Example outputs by the best finetuned sequence-to-sequence model (BART) and the best few-shot prompted LLM (Gemma-2) on the first example of the validation set, together with their evaluation scores on this specific example.