ULTRA: Unleash LLMs' Potential for Event Argument Extraction through Hierarchical Modeling and Pair-wise Self-Refinement

Xinliang Frederick Zhang[1,*], **Carter Blum**[2],
Temma Choji[2], **Shalin Shah**[2], and **Alakananda Vempala**[2]

[1]Computer Science and Engineering, University of Michigan
[2]Bloomberg
[1]xlfzhang@umich.edu
[2]{cblum18, tchoji, sshah804, avempala}@bloomberg.net

Abstract

Structural extraction of events within discourse is critical since it avails a deeper understanding of communication patterns and behavior trends. Event argument extraction (EAE), at the core of event-centric understanding, is the task of identifying role-specific text spans (i.e., *arguments*) for a given event. Document-level EAE (DocEAE) focuses on arguments that are scattered across an entire document. In this work, we explore open-source Large Language Models (LLMs) for DocEAE, and propose ULTRA, a hierarchical framework that extracts event arguments more cost-effectively. Further, it alleviates the *positional bias* issue intrinsic to LLMs. ULTRA sequentially reads text chunks of a document to generate a candidate argument set, upon which non-pertinent candidates are dropped through self-refinement. We introduce LEAFER to address the challenge LLMs face in locating the exact boundary of an argument. ULTRA outperforms strong baselines, including strong supervised models and ChatGPT, by 9.8% when evaluated by Exact Match (EM).

1 Introduction

Event extraction, a long-standing and prominent information extraction task, aims to extract event structures consisting of core information elements (e.g., "who" did "what" to "whom", "when", "where", and "why") from unstructured texts (Mourelatos, 1978; Riloff, 1996; Walker et al., 2005). Event-centric understanding is of great importance, not only in its inherent merits, but also due to its role as an information-rich representation for downstream tasks like summarization (Marujo et al., 2017; Li et al., 2021a), recommendation (Lu et al., 2016; Li et al., 2020a), and news narrative understanding (Jin et al., 2022; Zhang et al., 2022; Keith Norambuena et al., 2023). Event argument extraction (EAE), a crucial and challenging step in Event Extraction, is the task of identifying role-specific text spans (i.e., *arguments*) for a given event (Nguyen et al., 2016; Kar et al., 2020).

News title: Drought puts 2.1 million Kenyans at risk of starvation
News body:
[0] National disaster declared as crops fail after poor rains and locusts, while ethnic conflicts add to crisis Last modified on Wed 15 Sep 2021 07.02 BST.
[1] An estimated 2.1 million Kenyans face starvation due to a drought in half the country, which is affecting harvests.
[2] The National Drought Management Authority (NDMA) said people living in 23 counties across the arid north, northeastern and coastal parts of the country will be in "urgent need" of food aid over the next six months, after poor rains between March and May this year .
[3] The crisis has been compounded by Covid-19 and previous poor rains, it said, predicting the situation will get worse by the end of the year, as October to December rains are expected to be below normal levels.
. . .
[6] In July, the UN Food and Agriculture Organization in Kenya said the country needed 9.4bn Kenyan shillings (£62m) to mitigate the effects of the drought between July and November. · · ·
Event type: Droughts **Argument role:** Date

Baseline model outputs:
Flan-UL2: Wed 15 Sep 2021 **ChatGPT:** Wed 15 Sep 2021

ULTRA **outputs**
Layer-1 only: {March and May, July, Wed 15 Sep 2021}
Layer-1 + LEAFER**:** { between March and May this year , July, Wed 15 Sep 2021}
Full model: { between March and May this year , Wed 15 Sep 2021}

Table 1: Sample example from DocEE dataset, and outputs of select baselines and ULTRA. The ground-truth span is between March and May this year . ULTRA can correct itself with the LEAFER module, and drops less-pertinent candidates like "July". In contrast, both Flan-UL2 and ChatGPT fail to extract since sentence [0] includes a strong distractor, *"Wed 15 Sep 2021"*.

Existing EAE research mainly focuses on sentence-level understanding (Chen et al., 2015; Du and Cardie, 2020b; Lu et al., 2021) on the prevalent ACE 2005 dataset (Walker et al., 2005). Yet, in news, events are usually described at the document level, and arguments are typically scattered across an entire article (Hamborg et al., 2019). Thus, there is a pressing need to systematically study the document-level EAE (DocEAE) task, since sentence-level EAE systems fail to accommodate long-distance dependency (Ebner et al., 2020), cross-sentence inference (Li et al., 2021b) and multi-answer (Tong et al., 2022) problems intrinsic to DocEAE. Traditional supervised approaches consume large-scale annotations (e.g., Zheng et al., 2019; Pouran Ben Veyseh et al., 2022, more than 30,000 annotated articles required) in order to

[*] Work done during XFZ's internship at Bloomberg AI.

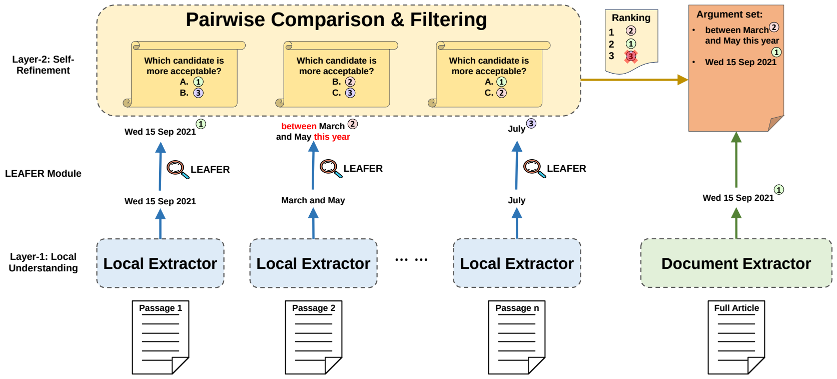

Figure 1: The overall architecture of ULTRA+, which consists of ULTRA (*left part*) and a document extractor (*bottom right*). In ULTRA, local extractors (layer-1) first generate a candidate argument set by comprehending text chunks sequentially, upon which *self-refinement* (layer-2) is performed through pairwise comparison to filter out less pertinent candidates. The predicted boundaries in the initial candidate set are rectified by the LEAFER module.

excel, and the state-of-the-art EAE model requires manual designs of templates for each argument role (Hsu et al., 2022). These approaches are not only costly but also not generalizable, since they cannot handle emerging events (Yang et al., 2023).[1] Recently, there has been a notable surge in applications of Large Language Models (LLMs) for NLP tasks, especially closed models such as Claude (Bai et al., 2022), PaLM (Chowdhery et al., 2022) and GPT-4 (OpenAI, 2023). The most relevant works to ours are Li et al. (2023); Han et al. (2023), but they only perform preliminary analysis by assessing ChatGPT's capability of solving IE tasks. Meanwhile, there is no prior research that has attempted to leverage LLMs to tackle DocEAE. In our preliminary investigation, we identified at least three challenges that arise when employing closed LLMs: 1) hitting endpoints incurs substantial costs and poses scalability challenges at inference; 2) undesirable prompt hacking is needed to ensure performance (Ouyang et al., 2022); 3) given the nature of news, where information is spread across the article, LLMs suffer from the *positional bias issue* (a.k.a, *lost in the middle;* Hou et al., 2024; Liu et al., 2024). Detailed literature review is in Appendix A.

To this end, we propose an easy-to-use framework that **U**nleashes **L**LMs' potential for event argument ex**TRA**ction through hierarchical modeling and pair-wise refinement, dubbed ULTRA. ULTRA, built on Flan-UL2 (Tay et al., 2022), first sequentially reads text chunks of a news article to generate a candidate argument set. ULTRA then learns to drop less-pertinent candidates through *self-refinement* by means of pairwise comparison. The LEAFER module, **LEA**rning **F**rom **ER**rors, is implemented to improve boundary identification of an argument span. Finally, we augment ULTRA with a document-level extractor to capture arguments that require reasoning of the full article.

Results on DocEE benchmark (Tong et al., 2022) show that ULTRA outperforms strong baselines, e.g., previous state-of-the-art supervised models and ChatGPT, by at least 9.8% and 7.5% when evaluated by the Exact Match (EM) and Head Noun Phrase Match (HM) metrics, but at a considerably reduced monetary cost. Existing methods only cater to improving precision, while our ULTRA significantly boosts recall as well (39.4 EM vs. 25.2). Besides better performance and lower costs, ULTRA also doesn't require specialized prompts, alleviates the *positional bias* issue, and grants stronger generalizability.

2 Methodology

Taking as input a news article $\mathbf{d}$, ULTRA first reads text chunks of the article $\mathbf{d}$ sequentially to generate a candidate argument set $\{\mathbf{a}\}$ (§2.1), upon which ULTRA drops less-pertinent candidates through self-refinement and returns $\{\mathbf{a}^f\}$ (§2.3). A LEAFER module, **LEA**rning **F**rom **ER**rors (§2.2), is introduced to tackle LLMs' incapability of locating exact boundaries of argument spans, and yield $\{\mathbf{a}'\}$. ULTRA+ is a variant augmented with extractions by a document-level extractor to capture information that requires full-article discourse analysis (e.g., extracting "why"-type arguments; §2.4). Figure 1 depicts the overall framework of ULTRA+.

Putting it all together, we produce two versions: ULTRA-base and ULTRA-long, which consume 5-sentence and 15-sentence windows in layer-1, respectively. Instead of conducting costly prompt hacking (Ouyang et al., 2022), we adopt an existing instruction from NIv2 (Wang et al., 2022) and tailor

[1] COVID-19 became an emerging topic since 2020 (Wang et al., 2020; Zhang et al., 2021), but not covered in traditional EE corpora (Walker et al., 2005; Ebner et al., 2020).

it to our use case, named **aligned instruction**. We show designed task instructions $\{\mathbf{I}\}$ in Table A4.

2.1 Layer-1: Local Understanding

Given a document $\mathbf{d}$, we first divide $\mathbf{d}$ to multiple k-sentence passage windows with a step size of $\lfloor \frac{k}{2} \rfloor$, denoted as $\{w_1, w_2, \cdots, w_l\}$. We adopt a fixed-window-size approach instead of a fixed-sequence-length approach (Devlin et al., 2019; Sun et al., 2019; Pappagari et al., 2019), which might cut a sentence in the middle, to allow each local extractor to comprehend each passage window in its entirety. Instantiated with Flan-UL2, the **local extractor** takes as input the concatenation of a task instruction ($\mathbf{I}$), a passage window (w_i), and a question written in natural language (q_j), e.g., *What is the "date" for the "Tsunami" event?* We prompt the local extractor in a zero-shot fashion[2] and explicitly instruct it to generate *N/A* if the input passage does not contain any relevant answer. After deduplication, we end up with a candidate argument set $\{\mathbf{a}\}_j$ for each question q_j.[3]

2.2 LEAFER Module

LLMs are deemed to have a knack for extracting relevant information (Li et al., 2023; Han et al., 2023), but we notice that LLMs still suffer from pinpointing the exact boundary of an argument span. Specifically, as shown in Figure 1, local extractions ($\{\mathbf{a}\}$) contain an apparently sensible answer "March and May" to the question "What is the 'Date' for the 'Droughts' event?", which is *lexically similar but semantically different* from the ground-truth answer "between March and May this year". To this end, we introduce a new module, LEAFER, short for ***LEA****rning* ***F****rom* ***ER****rors*, to alleviate this issue. The LEAFER module is a small-scale LM trained on errors produced by Flan-UL2. The trained LEAFER is employed to generate a **judgment**, which is to explicitly inform what is wrong and why it is wrong. The insightful judgment enables ULTRA to rectify boundaries of candidate arguments in $\{\mathbf{a}\}$ and produce $\{\mathbf{a}'\}$.

To support the training of LEAFER, we construct a LEAFER Bank using the few-shot training set of 50 annotated articles. Specifically, we prompt the same layer-1 local extractor to extract arguments for each (text chunk, question) input pair using the approach outlined in §2.1. For each input pair, we match the machine-extracted argument span with the corresponding ground-truth answer to produce a **judgment** automatically. Then, LEAFER is fine-tuned on this LEAFER Bank and trained to generate a judgment given an input pair and the machine-extracted answer. In this study, we instantiate the LEAFER module with Flan-T5-large.

2.3 Layer-2: Self-Refinement

While LEAFER addresses the semantic drift and imprecise boundary issues, ULTRA exhibits an *over-generation* issue due to window-based local extractors. Seeing recent success leveraging an LLM as a judge (Zheng et al., 2023; Wang et al., 2023a), we propose a **self-refinement** module that allows ULTRA to reflect on candidate arguments ($\{\mathbf{a}'\}$), and drop less pertinent candidates through pairwise ranking. There are usually two variations of LLM-as-a-judge: single-answer grading and pairwise comparison. As studied in Zheng et al. (2023) and observed in our preliminary study, we find that single answer grading cannot serve as an effective refinement judge since 1) absolute scores are extremely inflated and a considerably large portion of scores are close to 1 on a scale of 0 to 1; and 2) single answer grading fails to capture subtle differences between a specific pair. Therefore, in layer-2, we leverage **ranking by pairwise comparison** (Jamieson and Nowak, 2011; Lee and Vajjala, 2022; Jiang et al., 2023) to obtain the final argument set, $\{\mathbf{a}^f\}$, by first prompting Flan-UL2 to pick a better answer between a candidate pair, then ranking all candidates by aggregating pairwise-comparison scores, and finally filtering out candidates at low positions. To support dynamic filtering, we decide on $|\{\mathbf{a}^f\}|$ as follows:

$$|\{\mathbf{a}^f\}| = \lfloor 1 + \log_2(|\{\mathbf{a}'\}|) \rfloor \qquad (1)$$

The pairwise comparison produces a non-trivial score and catches nuanced differences, but is still trapped by the *positional bias* (Ko et al., 2020; Wang et al., 2023a; Liu et al., 2024) and *lack of scalability* due to the quadratic growth in pairwise comparisons. To mitigate these two issues, we resort to calibration and pruning, respectively.

Calibration. In layer-2, positional bias refers to a model's tendency to assign a higher score to an option at a particular position in a list, which has been shown to exist in ChatGPT/GPT-4 (Wang et al., 2023a). The issue is manifested as Flan-UL2 biasing towards an earlier displayed candidate. Drawing on the *Contextual Calibration* (Zhao et al., 2021), as demonstrated in eq. (2), we calibrate the

[2] We observe that few-shot prompting yields inferior results

[3] For brevity, we omit the subscript j in main contents.

raw probabilities of each option between a pair to reveal the truthful probabilities, i.e., $\mathrm{P}(\mathbf{a_i}|\mathbf{d})$.

$$\mathrm{P}(\mathbf{a_i}|\mathbf{d}) = \mathrm{softmax}(\mathbf{g}(\mathrm{P}(\mathbf{a_i}|\mathbf{d},\mathbf{I};\theta),\mathrm{P}(\mathbf{a_i}|\mathbf{I};\theta))) \tag{2}$$

where $\mathrm{P}(\mathbf{a_i}|\cdot)$ denotes the probability of an argument $\mathbf{a_i}$ being preferred given a certain input, and $\mathbf{d}$ and $\mathbf{I}$ denote the article and task instruction (see Table A4 for the instruction). Following Zhao et al. (2021), $\mathbf{g}(\mathbf{x},\mathbf{y})$ is a calibration function that can be instantiated as additive, $\mathbf{g}(\mathbf{x},\mathbf{y}) = \mathbf{x}-\mathbf{y}$, or multiplicative functions, $\mathbf{g}(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}}{\mathbf{y}}$. Using our designed comparison instruction ($\mathbf{I}$), we compute the prior probability $\mathrm{P}(\mathbf{a_i}|\mathbf{I};\theta)$ by leaving the $\{article\}$ field blank, while we fill in a concrete article when computing raw probability $\mathrm{P}(\mathbf{a_i}|\mathbf{d},\mathbf{I};\theta)$. With this calibration, we manage to alleviate the *positional bias* induced by the input template $\mathbf{I}$ and the innate bias of LLMs, θ.

Pruning. To tackle the scalability issue in which the number of comparisons grows quadratically, we prune the candidate set $\{\mathbf{a}'\}$ upfront to shrink its size. Specifically, we design a strategy that aligns with the fundamental principles of news journalism, wherein journalists prioritize the presentation of crucial information at the outset of a news story, known as the "inverted pyramid" structure (Pottker, 2003; Hamborg et al., 2019; Liu et al., 2022). That is, we consider up to 5 *earliest* candidate arguments, where the earliness of an argument is determined by its first occurrence in a news article. Our pruning strategy empirically reduces the number of subsequent pairwise computations by half. We also find that pruning itself can improve precision, even without pairwise comparisons. This further illuminates the validity of our designed pruning strategy.

2.4 Ensembling: ULTRA+

The ensembling technique consistently improves performance for a wide array of NLP tasks (Wang et al., 2019; Ganaie et al., 2022; Pitis et al., 2023; Jiang et al., 2023). LLM-Blender attempts to ensemble various LLMs on output space (Jiang et al., 2023), which prohibitively demands many computational resources. Instead, we suggest a simpler and more efficient approach: merging outputs by both ULTRA and a document-level argument extractor, which reads the full article and a question when extracting arguments. This way, we manage to combine the benefits of both local (*high recall*) and document-level (*high precision*) extractions.

Similar to Labrak et al. (2023); Han et al. (2023), we also observe marginal improvement on the dev set when providing in-context examples. To reduce inference-time overhead, we prompt the document-level extractor in a zero-shot manner.

3 Experiments

We conduct experiments on the DocEE dataset (Tong et al., 2022), which contains 27,485 news articles, classified into 59 event types, and 356 argument roles. We use the cross-domain setting in our experiments since it only contains a minimally annotated target training set (i.e., 50 articles) which can best assess models' generalizability in the wild. Specifically, its test set contains $1,955$ news articles covering 10 different event types, and each article is annotated with ~ 6.5 arguments. We use the same data split and processed texts as in the original DocEAE dataset for a fair comparison.

In terms of evaluation metrics, we follow the literature on document-level event argument extraction (Du and Cardie, 2020a; Tong et al., 2022), and adopt Exact Match (EM) and Head Noun Phrase Match (HM) as evaluation metrics. EM assesses if an extracted argument exactly matches a reference, while HM is a relaxed metric that concerns if there is an overlap of head words of noun phrases between extractions and references.

Baselines. We compare ULTRA against three model families to comprehensively study extraction performance and monetary costs. The first model family, *Supervised ML*, characterized as using human annotations as the supervision signal to train small-scale LMs, consists of EEQA (Du and Cardie, 2020b) and Ontology QA (Tong et al., 2022). Ontology QA, an extension of EEQA, incorporates argument ontology knowledge, which achieves the SOTA performance for DocEAE. Second, we compare with ChatGPT using different prompting techniques, given its popularity and impressive capability. We follow Li et al. (2023) and prompt ChatGPT to extract spans for *all argument roles* in one pass. For single-question variant, we modify the original prompt to instruct ChatGPT to extract span(s) for *only one argument role* at a time. Motivated by Wang et al. (2023b), which generates a chain-of-thought rationale before summarizing an article, we build the CoT-ChatGPT variant by replacing the summarizer in Wang et al. (2023b) with an argument extractor. The last family involves prompting a document-level extractor

Category	Method	Performance						Cost	
		EM			HM			Training	Inference
		P	R	F1	P	R	F1		
Supervised ML	EEQA* (Du and Cardie, 2020b)	29.4	20.3	24.0	68.1	46.9	55.5	$$$	∼0
	Onology QA* (Tong et al., 2022)	**36.6**	25.2	29.8	69.7	48.0	56.9		
Closed LLM	ChatGPT (Li et al., 2023)	35.6	18.0	23.9	74.4	58.0	65.2	0	$-$$
	ChatGPT (single question)	30.9	22.7	26.2	63.5	65.3	64.4		
	CoT-ChatGPT (Wang et al., 2023b)	31.2	16.2	21.3	71.0	55.2	62.1		
Flan-UL2	Custom instructions**	27.6	17.8	21.6	69.2	45.2	54.6	$	∼0
	Aligned instruction	36.1	20.7	26.3	**76.6**	52.0	62.0		
ULTRA (Ours)	ULTRA-base	29.0	34.5	31.5	61.8	70.3	65.8	$	∼0
	+ Ensemble (i.e., ULTRA+)	28.0	**39.4**	**32.7**	63.7	**75.3**	69.0		
	ULTRA-long	32.3	30.5	31.4	68.4	65.9	67.1		
	+ Ensemble (i.e., ULTRA+)	30.2	35.5	32.6	68.6	71.5	**70.1**		

Table 2: Results on DocEE dataset for document-level event argument extraction, and breakdown of EM and HM scores by precision (P), recall (R) and F1. We also report estimated monetary cost by model category, divided into training and inference costs (Appendix B). Best results are **bold**. ULTRA achieves the best F1 performances at a reduced cost, and reaches higher recalls than any baseline. We use the additive function when performing calibration (§2.3), though we see a trivial difference between additive and multiplicative functions. The ablation study results of ULTRA can be found in Table A2. *Results are taken from Tong et al. (2022). **Average results of 5 instructions, results of individual custom instructions are included in Table A3.

with different instructions, utilizing Flan-UL2 as its backbone for a fair comparison. This serves two purposes: test sensitivity to different custom instructions that are designed from scratch; and illuminate the effectiveness of *Aligned Instructions*.

4 Results and Analysis

Our proposed ULTRA achieves the best F1 scores across the board (Table 2), especially compared to two strong baseline families, *Supervised ML* and *Closed LLM*, at a considerably reduced monetary cost in training and inference. ULTRA also significantly improves the EM recall by 56% over the best-performing baseline model (39.4 vs. 25.2), demonstrating robust generalizability considering ULTRA's exposure to at most 5-shots per event type.

Seeing the relatively low EM scores, we explore the errors and include case analyses in Table A1. We provide cost estimation in Appendix B.

Using ChatGPT for DocEAE. Despite the common flaw of outputting seemingly coherent assertions that are false in reality, known as hallucination (Manakul et al., 2023; Feldman et al., 2023), we recognize another issue, which seems to be less studied in the NLP community, that answer spans extracted by ChatGPT are *verbose* (Zheng et al., 2023; Chen et al., 2023a). This explains the reason why ChatGPT achieves the best HM scores in the literature since longer generations are more likely to contain relevant information, while the EM is low due to the *nature of verbosity*.

Further Study on Window Size. Despite ULTRA-base and ULTRA-long achieving almost identical EM F1 scores, they present different extraction properties, wherein ULTRA-base reaches the highest recall while ULTRA-long is more balanced. In this subsection, we specifically look into the extraction property of the Layer-1-only variant of ULTRA. Figure A1 shows the performance trend with the window size. We notice that precision steadily goes up while recall consistently goes down by increasing the window size. We attribute this trend to the fact that a larger window size leads to fewer text chunks being fed into ULTRA. It is also worth mentioning that the overall F1 performance plateaus after a window size of 15. This observation underscores a key aspect of ULTRA: its flexibility for accommodating various extraction criteria. For instance, when the objective is to harvest the most relevant information, opting for a smaller window size appears to be a favorable choice. Conversely, selecting a larger window is advisable if precision is the core of a product or the target audience consists of vulnerable populations susceptible to misinformation.

5 Conclusion

In this study, we present ULTRA, a cost-effective event argument extraction framework built upon an open-source LLM. Concretely, ULTRA reads a sequence of text chunks from an article, the outputs of which are refined through self-refinement. With minimal annotation efforts, a LEAFER module is implemented to improve argument span boundary identification. Our results show the superiority of ULTRA in comparison to *supervised ML* and *closed LLMs*. We further showcase the customizability of ULTRA to cater to different extraction criteria.

Limitations

GPU resources. Despite ULTRA managing to reduce monetary cost, it still requires advanced computational resources. Specifically, we deploy ULTRA on a single NVIDIA A100 (80GB) with significant CPU and memory resources. Due to budget constraints, we truncate an input if it is longer than $2,048$ tokens.

DocEE benchmark. Due to the limited large-scale, well-regarded datasets for the DocEAE task, we only conduct experiments on one benchmark – DocEE dataset (Tong et al., 2022). DocEE, though covering 59 event types and 356 argument roles, is still not comprehensive. Therefore, our results might not truthfully reveal the generalizability of ULTRA. In future work, we will explore how to examine the true generalizability of developed systems in the wild.

Acknowledgements

The authors thank the Bloomberg AI team for their support of this work. We would like to thank Shijie Wu for hardware support and discussions on FLAN-UL2 setup configurations. We appreciate Igor Malioutov, Daniel Preoţiuc-Pietro and Umut Topkara for their insightful inputs and helpful feedback at different stages of this project. We also thank ACL ARR reviewers for their helpful feedback.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, John Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, E Perez, Jamie Kerr, Jared Mueller, Jeff Ladish, J Landau, Kamal Ndousse, Kamilė Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noem'i Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, T. J. Henighan, Tristan Hume, Sam Bowman, Zac Hatfield-Dodds, Benjamin Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom B. Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *ArXiv*, abs/2212.08073.

Lingjiao Chen, Matei Zaharia, and James Y. Zou. 2023a. How is chatgpt's behavior changing over time? *ArXiv*, abs/2307.09009.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023b. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020a. Document-level event role filler extraction using multi-granularity contextualized encoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8010–8020, Online. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020b. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Philip G. Feldman, James R. Foulds, and Shimei Pan. 2023. Trapping llm hallucinations using tagged context prompts. *ArXiv*, abs/2306.06085.

M. A. Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N. Suganthan. 2022. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.*, 115:105151.

Jun Gao, Huan Zhao, Yice Zhang, Wei Wang, Changlong Yu, and Ruifeng Xu. 2023. Benchmarking large language models with augmented instructions for fine-grained information extraction. *ArXiv*, abs/2310.05092.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

Felix Hamborg, Corinna Breitinger, and Bela Gipp. 2019. Giveme5w1h: A universal system for extracting main events from news articles. In *INRA@RecSys*.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *ArXiv*, abs/2305.14450.

Xinyu He, Ping Tai, Hongbin Lu, Xin Huang, and Yonggong Ren. 2022. A biomedical event extraction method based on fine-grained and attention mechanism. *BMC Bioinformatics*, 23.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. 2022. Training compute-optimal large language models. *ArXiv*, abs/2203.15556.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *ECIR*.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

Kevin G. Jamieson and Robert D. Nowak. 2011. Active ranking using pairwise comparisons. *ArXiv*, abs/1109.3701.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Xiaomeng Jin, Manling Li, and Heng Ji. 2022. Event schema induction with double graph autoencoders. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2013–2025, Seattle, United States. Association for Computational Linguistics.

Debanjana Kar, Sudeshna Sarkar, and Pawan Goyal. 2020. Event argument extraction using causal knowledge structures. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 287–296, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. 2023. A survey on event-based news narrative extraction. *ACM Comput. Surv.*, 55(14s).

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. 2023. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. *ArXiv*, abs/2307.12114.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *ArXiv*, abs/2304.11633.

Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021a. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6443–6456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman. 2020a. GAIA: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, Online. Association for Computational Linguistics.

Qingquan Li, Qifan Zhang, Junjie Yao, and Yingjie Zhang. 2020b. Event extraction for criminal legal text. *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 573–580.

Sha Li, Heng Ji, and Jiawei Han. 2021b. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yujian Liu, Xinliang Zhang, Kaijian Zou, Ruihong Huang, Nicholas Beauchamp, and Lu Wang. 2023. All things considered: Detecting partisan events from news media with cross-article comparison. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15472–15488, Singapore. Association for Computational Linguistics.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Di Lu, Clare Voss, Fangbo Tao, Xiang Ren, Rachel Guan, Rostyslav Korolov, Tongtao Zhang, Dongang Wang, Hongzhi Li, Taylor Cassidy, Heng Ji, Shih-fu Chang, Jiawei Han, William Wallace, James Hendler, Mei Si, and Lance Kaplan. 2016. Cross-media event extraction and recommendation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 72–76, San Diego, California. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. 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 2795–2806, Online. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Luís Marujo, Ricardo Ribeiro, Anatole Gershman, David Martins de Matos, Joao P. Neto, and Jaime G. Carbonell. 2017. Event-based summarization using a centrality-as-relevance model. *Knowledge and Information Systems*, 50:945–968.

Alexander P. D. Mourelatos. 1978. Events, processes, and states. *Linguistics and Philosophy*, 2:415–434.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Liangming Pan, Michael Stephen Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *ArXiv*, abs/2308.03188.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021*.

R. Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844.

Silviu Pitis, Michael Ruogu Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *ArXiv*, abs/2304.05970.

Horst Pottker. 2003. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022. MEE: A novel multilingual event extraction dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *AAAI/IAAI, Vol. 2*.

Omar Sharif, Madhusudan Basak, Tanzia Parvin, Ava Scharfstein, Alphonso Bradham, Jacob T. Borodovsky, Sarah E. Lord, and Sarah Masud Preum. 2024. Characterizing information seeking events in health-related social discourse. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 22350–22358. AAAI Press.

Shirong Shen, Guilin Qi, Zhen Li, Sheng Bi, and Lusheng Wang. 2020. Hierarchical Chinese legal event extraction via pedal attention mechanism. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 100–113, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Saurabh Srivastava, Gaurav Singh, Shou Matsumoto, Ali K. Raz, Paulo C. G. Costa, Joshua C. Poore, and Ziyu Yao. 2023. Mailex: Email event and argument extraction. *ArXiv*, abs/2305.13469.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*.

Beth M. Sundheim. 1992. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier García, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2022. Ul2: Unifying language learning paradigms. In *International Conference on Learning Representations*.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United States. Association for Computational Linguistics.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. Ace (automatic content extraction) english annotation guidelines for events. *Linguistic Data Consortium,*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *ArXiv*, abs/2305.17926.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *ArXiv*, abs/2302.10205.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3533–3546, Online. Association for Computational Linguistics.

Xianjun Yang, Yujie Lu, and Linda Petzold. 2023. Few-shot document-level event argument extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8029–8046, Toronto, Canada. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. 2022. Generative entity-to-entity stance detection with knowledge graph augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9950–9969, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinliang Frederick Zhang, Heming Sun, Xiang Yue, Simon Lin, and Huan Sun. 2021. COUGH: A challenge dataset and models for COVID-19 FAQ retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3759–3769, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinliang Frederick Zhang, Winston Wu, Nick Beauchamp, and Lu Wang. 2024. MOKA: Moral knowledge augmentation for moral event extraction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico. Association for Computational Linguistics.

Weizhong Zhao, Jinyong Zhang, Jincai Yang, Tingting He, Huifang Ma, and Zhixin Li. 2020. A novel joint biomedical event extraction framework via two-level modeling of documents. *Inf. Sci.*, 550:27–40.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 337–346, Hong Kong, China. Association for Computational Linguistics.

A Related Work

A.1 Event Argument Extraction (EAE)

Most event argument extraction research has experimented on the 2005 Automatic Content Extraction (ACE 2005; Walker et al., 2005), while recent work delves into domain-specific areas such as biomedical texts (Zhao et al., 2020; He et al., 2022), legal documents (Li et al., 2020b; Shen et al., 2020), partisan contents (Liu et al., 2023), morality-bearing contents (Zhang et al., 2024), and conversations (Srivastava et al., 2023).

Existing work primarily focused on the sentence-level event understanding task. Methods can be categorized under one of the three following approaches: sequence labeling (Chen et al., 2015; Nguyen et al., 2016) where Lin et al. (2020) further constrains the inference with global features; question answering (Du and Cardie, 2020b), which includes ontology knowledge about argument roles; and generative information extraction (Paolini et al., 2021; Lu et al., 2021). Particularly, DEGREE reformulates EAE as template-based conditional generation, and archives impressive performance on various benchmarks (Hsu et al., 2022). Yet, it demands huge annotation efforts, which require one template for each argument role, and is therefore not generalizable. In this work, we are seeking to improve EAE performance with general instructions instead of argument-specific templates.

Lately, there has been an increasing interest in document-level EAE (DocEAE), since events are usually described at the document level and arguments are usually scattered across multiple sentences (Sundheim, 1992; Hamborg et al., 2019; Tong et al., 2022). For example, RAMS (Ebner et al., 2020) and MEE (Pouran Ben Veyseh et al., 2022) both define “document” as a 5-sentence segment. In contrast, WikiEvents (Li et al., 2021b) and DocEE (Tong et al., 2022) present full articles and focus on argument extractions for the main event. In this work, we use DocEE as a benchmark since it features broad coverage of event types in the news domain. Methodology-wise, Du and Cardie (2020a) and Tong et al. (2022) handle DocEAE by extending sentence-level labeling and question-answering approaches, respectively. Li et al. (2021b) frames DocEAE as template-based conditional generation in the same vein as the sentence-level generative approach. Unfortunately, none of the aforementioned methods tackle the *argument-scattering* challenge; instead, they treat a full article as if it were an extended sentence. Zheng et al. (2019) is the first work to address this issue by modeling DocEAE as an entity-centric graph, which is further augmented with a “tracker” module to capture the interdependency among arguments and events (Xu et al., 2021). Nonetheless, the “tracker” is insufficient due to its limitation of not considering the results of later extractions when processing earlier ones. On the contrary, our `ULTRA` bridges the gap through the implementation of a *self-refinement* module, which is grounded in pairwise comparison and functions akin to a bi-directional tracker.

A.2 Using Large Language Models for IE

The past years have witnessed the rise of transformer architecture (Vaswani et al., 2017), paving the way for a series of powerful language models. Recent studies have evinced that scaling up model sizes yields more powerful abilities (Hoffmann et al., 2022), and unlocks an emergent ability that is not present in smaller models (Wei et al., 2022a). These large language models (LLMs), which often exceed a hundred billion parameters, are typically closed systems (i.e., no open checkpoints available). Notable examples include PaLM (Chowdhery et al., 2022), Claude (Bai et al., 2022), and GPT-4 (OpenAI, 2023). Numerous methods are also developed to enhance LLMs’ reasoning and problem-solving capabilities, such as chain-of-thought (Wei et al., 2022b), self-correction (Pan et al., 2023), and external tool (e.g., Python interpreter) augmentation (Gao et al., 2022; Chen et al., 2023b) among others.

ChatGPT,[4] one of the most burgeoning LLM, is trained on high-quality conversation datasets using reinforcement learning from human feedback (RLHF; Christiano et al., 2017), has led to a transformative wave. The most relevant to our research is leveraging ChatGPT for the information extraction task (Li et al., 2023; Han et al., 2023), including named entity recognition (Xie et al., 2023), temporal relation extraction (Yuan et al., 2023), event detection (Sharif et al., 2024), and event argument extraction (Wei et al., 2023). These papers’ primary focus is either benchmarking ChatGPT’s performance, which shows inferior results to specialized supervised IE systems (Li et al., 2023; Han et al., 2023), or curating new benchmark datasets (Gao et al., 2023). In contrast,

[4] `https://chat.openai.com/`

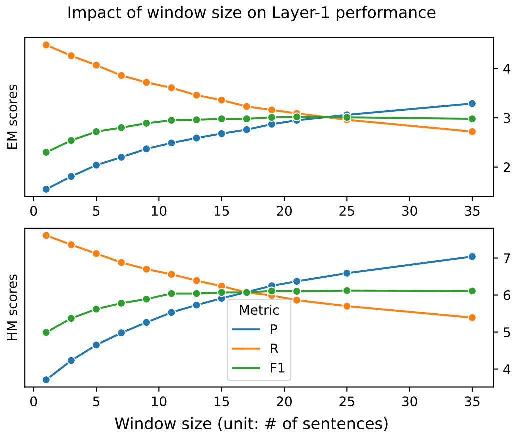

Figure A1: The impact of window size on the performance of the Layer-1-only variant of ULTRA. Results are based on dev set. With the window size, precision goes up while recall goes down since fewer chunks are fed into ULTRA. The F1 performances plateau after the window size of 15.

our proposed ULTRA framework outperforms strong baselines, including the previous state-of-the-art (SOTA) models, capitalizing on the effectiveness of our designed LEAFER and self-reflection modules. Besides, to the best of our knowledge, we are the first to exploit LLMs for the DocEAE task.

B Cost Estimation

In addition to models' extraction performance, Table 2 also presents the cost estimation of each model family. We briefly introduce the criteria used when estimating monetary costs. The training cost is mainly associated with document annotations.[6] Regarding inference, we consider the expenses incurred in hitting API endpoints.[7] Per Tong et al. (2022), both EEQA and Ontology QA are trained on 22K articles, each costs \$0.9, totaling \$20, 000. Based on ChatGPT pricing,[8] the base cost is \$0.004/1K tokens. Processing each article and then producing answers would consume 5K to 50K tokens on average, depending on the input mode. The test set contains 2K examples, so the total cost is around \$40 to \$400. For the Flan-UL2 baseline and our ULTRA, each only needs annotations of up to 50 articles, for the training of the LEAFER module. It is noteworthy that ULTRA enables cost-effective scaling at inference, while ChatGPT might face budget constraints.

News title: Experts: Oregon seems to be in ’perpetual drought’
News body:
[0] Experts say Oregon is becoming less resilient to drought as fewer seasons of abundant rain and snow prevent it from bouncing back from hot and dry conditions.
[1] Wheat at the farm of Nicole Berg in Washington’s Horse Heaven Hills shows signs of a drought in May 2021, with a damaged curled head.
[2] Anna King The Capital Press reports that Larry O’Neill, state climatologist at Oregon State University, says the current drought is “historically significant,” with about three-quarters of the state experiencing conditions considered “extreme” or “exceptional.”
[3] However, the state is actually in the fourth year of below-average precipitation, which has exacerbated the drought during “unprecedentedly” high temperatures this summer, O’Neill told the Oregon Water Resources Commission on Wednesday.
[4] “We don’t recover from droughts as quickly as we did previously,” he said. “We seem to be in perpetual drought.”
[5] Parched soils were insufficiently recharged with moisture over winter and spring, which has harmed vegetative growth, including crops and forage, said Ryan Andrews, a hydrologist at the Oregon Water Resources Department, which is overseen by the commission.
[6] Reservoir and stream flow levels are below average across most of the state, reducing water available to irrigators, while ranchers have sold off livestock due to poor rangeland conditions, he said.
[7] Fish die-offs followed the June heat wave in several important rivers basins, including the Willamette, Grande Ronde, John Day and along the North coast, Andrews said.
[8] The state would need plentiful rain and snow during the autumn to begin emerging from the drought, but the long-term federal climate forecast doesn’t anticipate such a reversal, he said.
[9] “We’re anticipating conditions to persist, at least in the near term.” Between March and July, the state received less rain than during any comparable period in nearly a century, O’Neill said.
[10] “The dry spring and summer is one of the main contributing factors to why this drought has become so severe.”
[11] An IBIS explores what habitat remains on the Klamath Basin’s wildlife refuges during a drought year that is exacerbating water resources challenges in this arid region.
[12] Devan Schwartz / OPB The area under “extreme” and “exceptional” drought ratings is the most extensive in Oregon since the start of the U.S. Drought Monitor more than 20 years ago, he said.
[13] The most severe “exceptional” level of drought now seen across one-fourth of the state would normally be expected to occur every 20 to 50 years, O’Neill said.
[14] However, droughts are judged by historical standards, so the concept of such “recurrence intervals” grows less valid as dry periods become more common, he said.
[15] “It’s going to take some time to get used to the new normal we’re experiencing right now,” O’Neill said.

Argument role 1: Related Rivers or Lakes
ULTRA: {“the Willamette, Grande Ronde, John Day and along the North coast”}
Ground-truth: {“Willamette, Grande Ronde, John Day and along the North coast”}

Argument role 2: Cause
ULTRA: {“fewer seasons of abundant rain and snow”}
Ground-truth: {“The dry spring and summer”, “fewer seasons of abundant rain and snow”}

Argument role 3: Areas Affected
ULTRA: {“Oregon”}
Ground-truth: {“Willamette, Grande Ronde, John Day and along the North coast”, “Washington”}

Table A1: Case analyses of ULTRA outputs for a Droughts event. For argument role 1, the system output is considered an exact match with the ground-truth annotation, since stop words are removed at evaluation time. For argument role 2, our ULTRA achieves 100% precision but 50% recall since the model fails to capture both reasons that caused the drought. For argument role 3, the system output is completely wrong.

[6] Here, we omit the sunk cost incurred due to pre-training.

[7] The server maintenance cost is considered low (~ 0).

[8] https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/

Model	Configuration	EM			HM		
		P	R	F1	P	R	F1
ULTRA-base	Layer-1 only	22.5	42.5	29.4	50.3	77.7	61.1
	Layer-1 + LEAFER	22.6	**43.4**	29.7	50.7	**78.4**	61.6
	Layer-1 + LEAFER + Layer-2	29.0	34.5	**31.5**	61.8	70.3	65.8
ULTRA-long	Layer-1 only	29.2	34.0	31.4	63.3	68.9	66.0
	Layer-1 + LEAFER	29.2	34.5	31.6	63.4	69.7	66.4
	Layer-1 + LEAFER + Layer-2	**32.3**	30.5	31.4	**68.4**	65.9	**67.1**

Table A2: Ablation study results of variants of ULTRA. ULTRA-base manages to improve recall by over-generating candidate answers, while the over-generation problem is redressed by self-refinement (layer-2) through pairwise comparison. ULTRA-long does not confront the over-generation issue, thus, layer-2 does not contribute significantly to the performance as in ULTRA-base. Best results are **bold**.

ID	EM			HM			Instruction template
	P	R	F1	P	R	F1	
0	26.5	18.9	22.0	67.6	45.2	54.2	The following is a news article about a “{e_type}”:\n{news}\nBy reading the above article, determine the “{arg_role}” for the “{e_type}”.
1	29.5	19.5	23.5	71.3	47.3	56.8	Here is a news article:\n{news}\nThe above news article is about a “{e_type}”. Identify “{arg_role}” for the “{e_type}” from the news article.
2	29.0	16.2	20.8	71.8	42.8	53.7	Read the following news article, and then answer questions. Context: {news} \nQuestion: Identify “{arg_role}” for this “{e_type}” event.
3	25.5	18.6	21.5	64.6	43.1	51.7	Given the following news about a “{e_type}”:\n{news}\nThe “{arg_role}” for the “{e_type}” is
4	27.6	15.7	20.0	70.5	47.5	56.8	Given the following news about a “{e_type}”:\n{news}\nWhat is the “{arg_role}” for the “{e_type}”?

Table A3: Performances of each individual custom instruction. {e_type}, {arg_role} and {news} are placeholders to be filled with event type, argument role and news content, respectively. Flan-UL2 is considerably sensitive to the input instruction, and even with a tiny change in the question, the model performance varies a lot, as manifested by contrasting instruction ID 3 and 4.

Stage	Instruction
Layer-1 local extractor	Given a passage from a news article about {e_type}, select the tokens representing information about ’{arg_role}’ or answer ’N/A’ if the question is not answerable. \nPassage: {sentence}. Question: What is the ’{arg_role}’ for the ’{e_type}’ event?
Layer-2 comparator	Given a news article about ’{e_type}’ and two candidate spans, decide whether ’{arg1}’ is a more acceptable ’{arg_role}’ than ’{arg2}’ for the ’{e_type}’ event. \nArticle: {article} \n For this ’{e_type}’ event, is ’{arg1}’ a more acceptable ’{arg_role}’ than ’{arg2}’? Answer yes/no.
Document-level extractor	Given a news article about {e_type}, select the tokens representing information about ’{arg_role}’. \n Context: {news}. Question: What is the ’{arg_role}’ for the ’{e_type}’ event?

Table A4: Instructions designed for each stage in ULTRA. The document-level extractor is utilized in the ensembling mode of ULTRA (§2.4), and serves as the Flan-UL2 baseline (§3). These *aligned instructions* are adapted from task 179 (participant extraction) in NIv2[5] (Wang et al., 2022).

Extraction	Ground Truth	Judgements
N/A	N/A	Yes.
something	N/A	No, you should generate “N/A”
N/A	something	No, you should generate “[GT]”.
something	something	Yes.
something	something longer OR something shorter	You are almost there! The right answer should be “[GT]”.
anything	something	No, you should generate “[GT]”.

Table A5: Designed template-based judgments used to train the LEAFER module in order to address the boundary identification issue. We categorize the (extraction, ground truth) pairs into six classes. Here, “anything” refers to a generated extraction that is completely off, and “[GT]” acts as a placeholder to be replaced with a specific ground-truth argument.