Bloomberg

Engineering

UNIVERSITY OF MICHIGAN

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]University of Michigan, [2]Bloomberg

ACL 2024

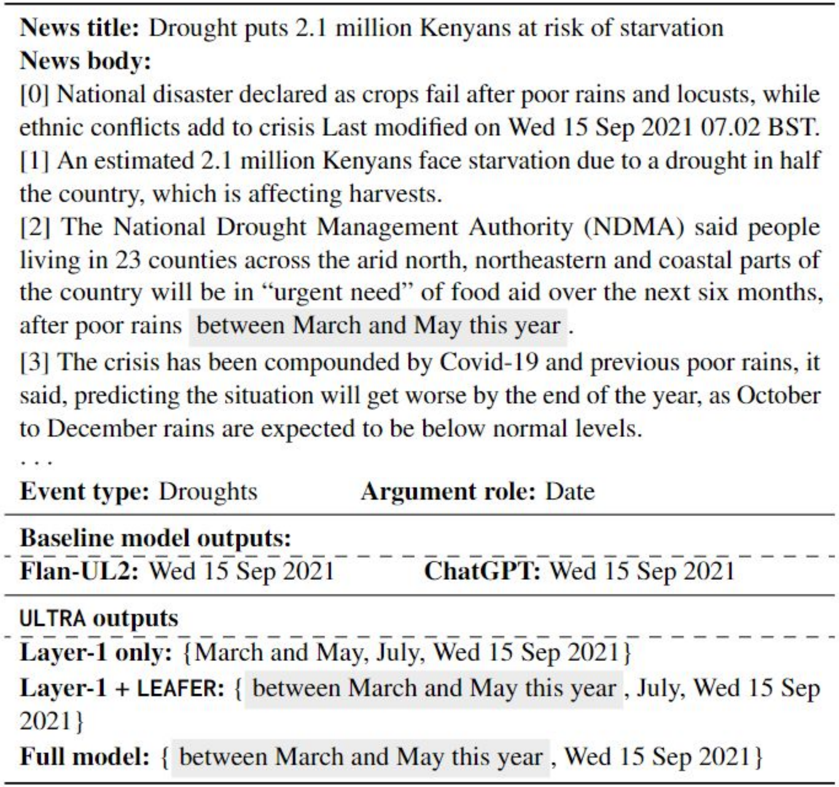
*Work done during XFZ's internship at Bloomberg AI

Document-level Event Argument Extraction (DocEAE)

Task Formulation: Provided with an input document of a particular event type, extract a set of phrases that mention an event-specific attribute (i.e., argument role).

Document-level Event Argument Extraction (DocEAE)

Task Formulation: Provided with an input document of a particular event type, extract a set of phrases that mention an event-specific attribute (i.e., argument role).



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.

. . .

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}

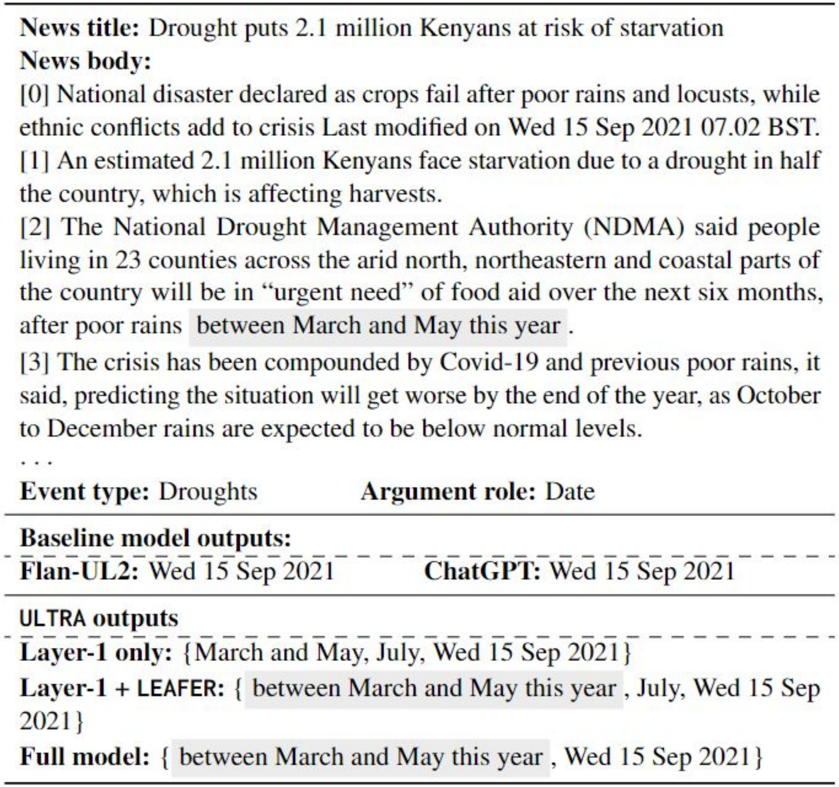
Full model: { between March and May this year , Wed 15 Sep 2021}

Document-level Event Argument Extraction (DocEAE)

Task Formulation: Provided with an input document of a particular event type, extract a set of phrases that mention an event-specific attribute (i.e., argument role).

Challenges of DocEAE:

- Long-distance Dependency
- Cross-sentence Inference (i.e., answers scattered across the document)
- Multi-answer (i.e., more than one plausible span for one argument role)



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.

. . .

Event type: Droughts **Argument role:** Date

Baseline model outputs:

Flan-UL2: Wed 15 Sep 2021 **ChatGPT:** Wed 15 Sep 2021

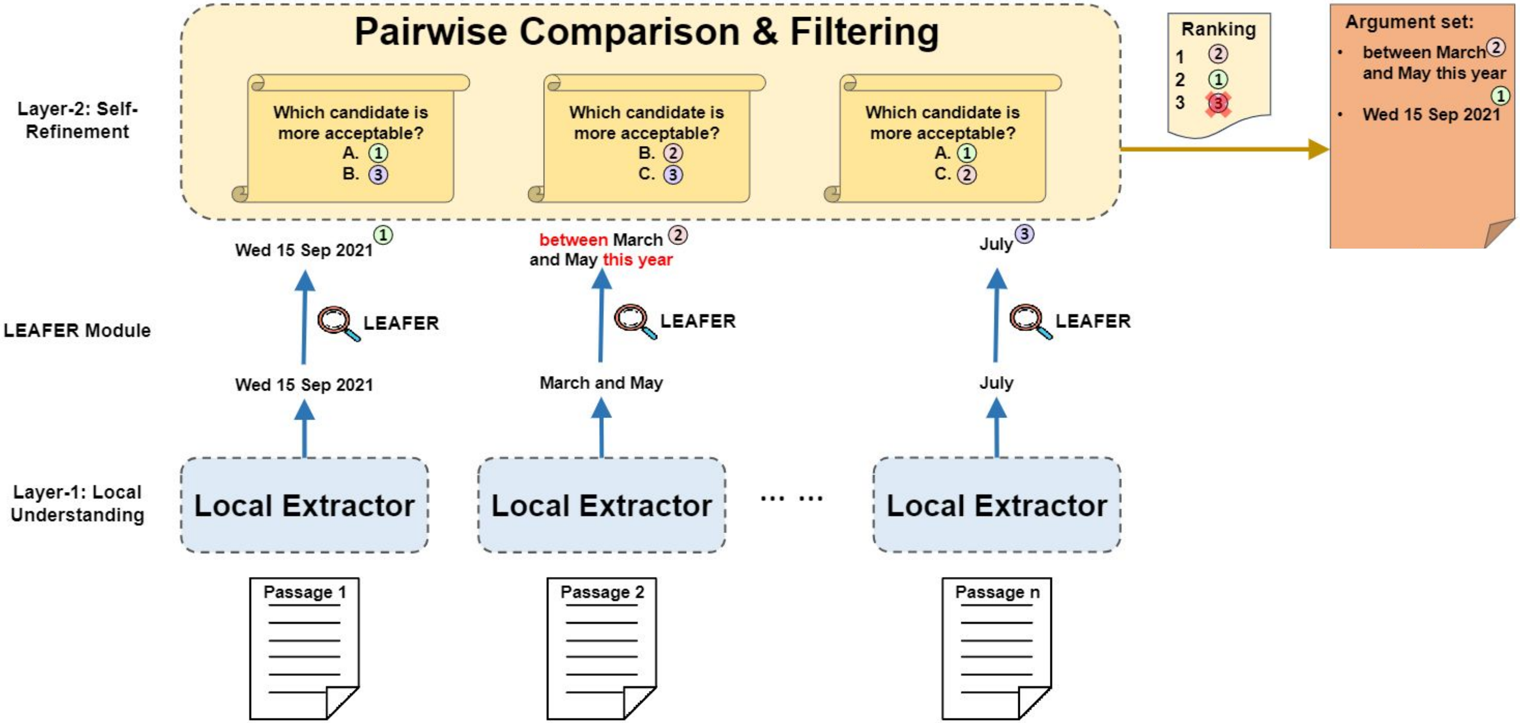
ULTRA outputs

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}

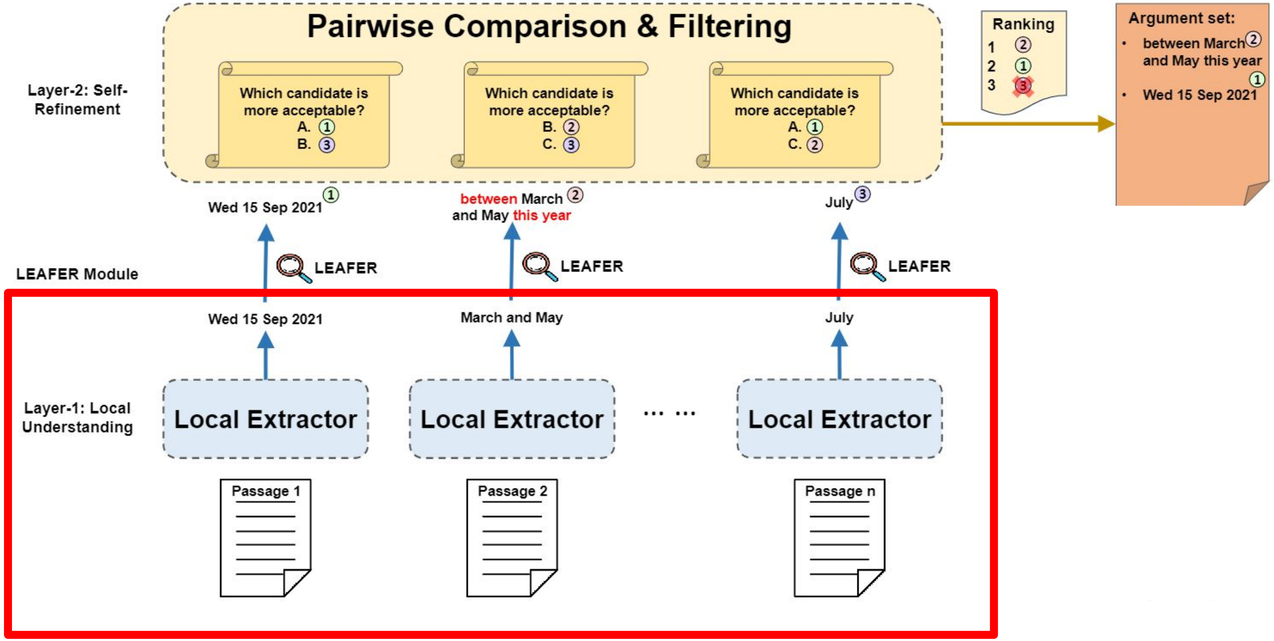
Full model: { between March and May this year , Wed 15 Sep 2021}

Model: ULTRA



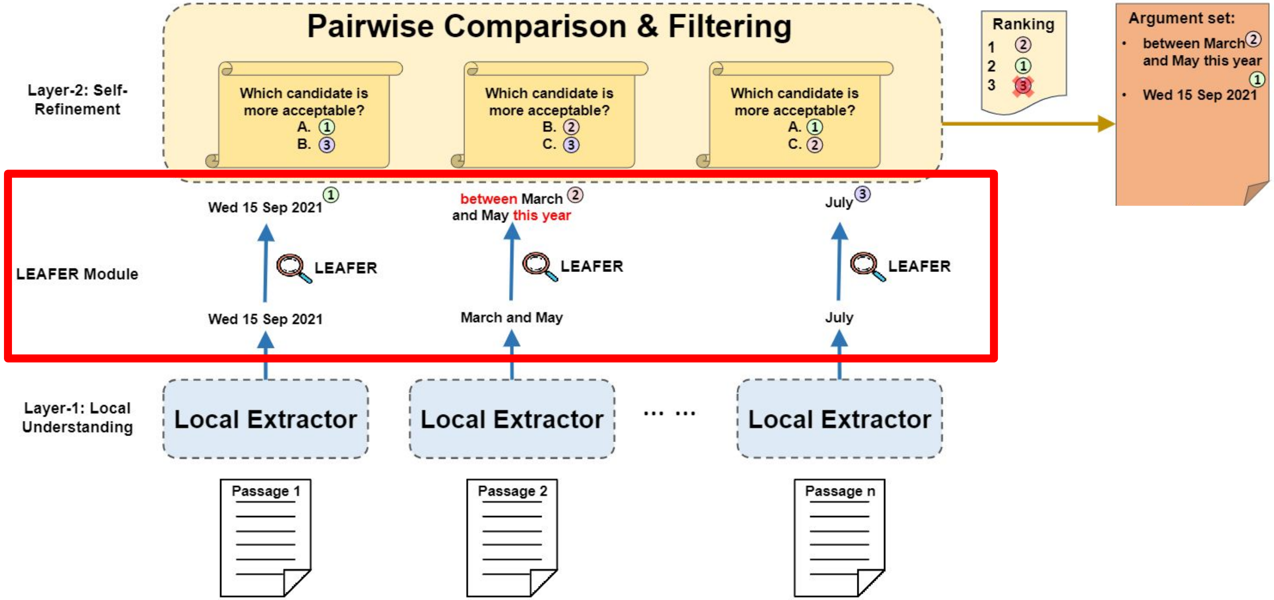
Pairwise Comparison & Filtering
Layer-2: Self-Refinement
Which candidate is more acceptable? A. ① B. ③
Which candidate is more acceptable? B. ② C. ③
Which candidate is more acceptable? A. ① C. ②
Ranking
1 ②
2 ①
3 ③
Argument set:
• between March ② and May this year
• Wed 15 Sep 2021 ①
Wed 15 Sep 2021 ①
between March and May this year ②
July ③
LEAFER Module
LEAFER
LEAFER
LEAFER
Wed 15 Sep 2021
March and May
July
Layer-1: Local Understanding
Local Extractor
Local Extractor
… …
Local Extractor
Passage 1
Passage 2
Passage n

ULTRA: Layer-1 Local Understanding


Pairwise Comparison & Filtering
Layer-2: Self-Refinement
Which candidate is more acceptable? A. ① B. ③
Which candidate is more acceptable? B. ② C. ③
Which candidate is more acceptable? A. ① C. ②
Ranking
1 ②
2 ①
3 ③
Argument set:
• between March ② and May this year
• ① Wed 15 Sep 2021
Wed 15 Sep 2021 ①
between March ② and May this year
July ③
LEAFER Module
LEAFER
LEAFER
LEAFER
Wed 15 Sep 2021
March and May
July
Layer-1: Local Understanding
Local Extractor
Local Extractor
… …
Local Extractor
Passage 1
Passage 2
Passage n

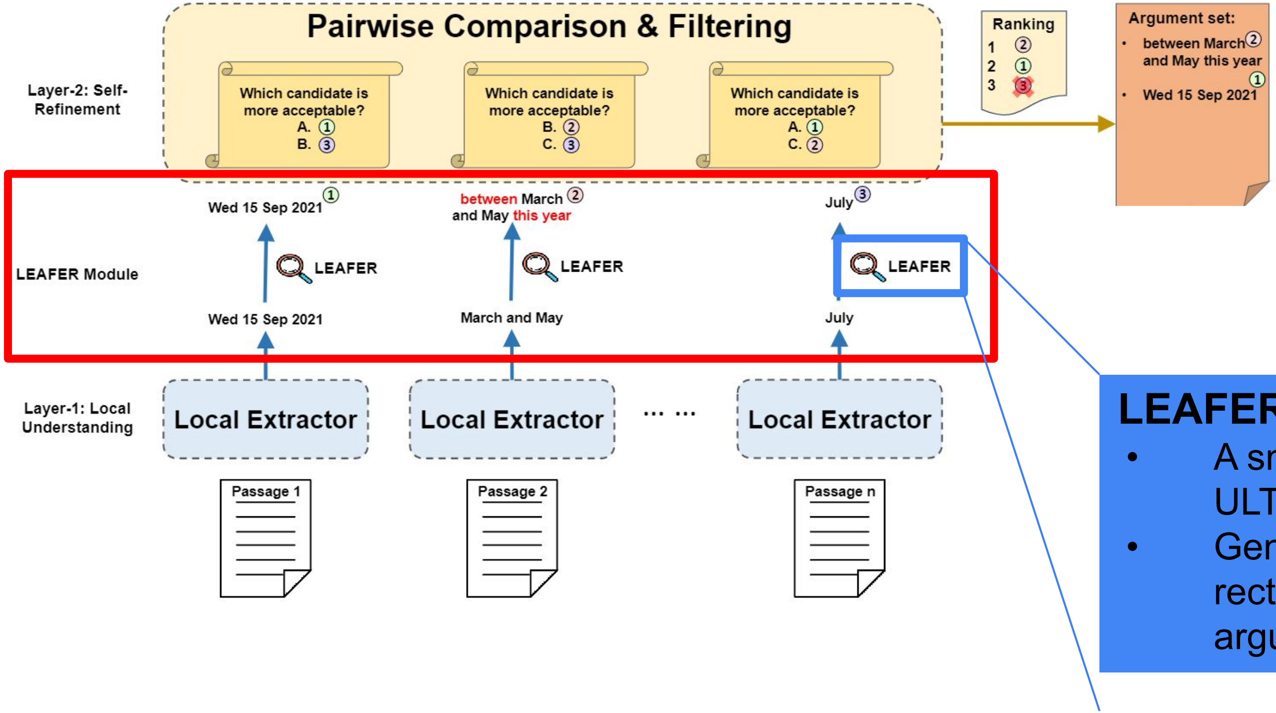
ULTRA first reads text chunks of an article sequentially to generate a candidate argument set **{a}**

ULTRA: LEAFER Module (Self-Reflection)


Pairwise Comparison & Filtering
Layer-2: Self-Refinement
Which candidate is more acceptable? A. ① B. ③
Which candidate is more acceptable? B. ② C. ③
Which candidate is more acceptable? A. ① C. ②
Ranking
1 ②
2 ①
3 ③
Argument set:
• between March ② and May this year
• Wed 15 Sep 2021 ①
LEAFER Module
Wed 15 Sep 2021 ①
between March ② and May this year
July ③
LEAFER
LEAFER
LEAFER
Wed 15 Sep 2021
March and May
July
Layer-1: Local Understanding
Local Extractor
Local Extractor
… …
Local Extractor
Passage 1
Passage 2
Passage n

A **LEAFER** module, LEArning From ERrors, is introduced to tackle LLMs' incapability of locating exact boundaries of arguments, and yield **{a'}.**

ULTRA: LEAFER Module (Self-Reflection)

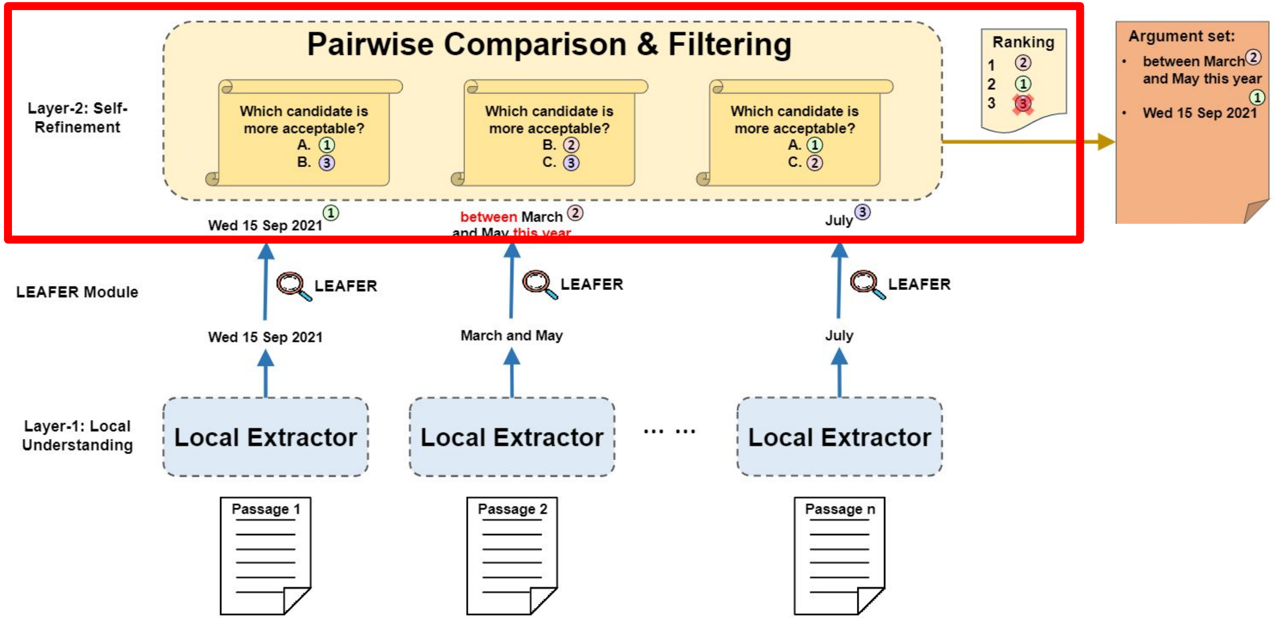

Pairwise Comparison & Filtering
Layer-2: Self-Refinement
Which candidate is more acceptable? A. ① B. ③
Which candidate is more acceptable? B. ② C. ③
Which candidate is more acceptable? A. ① C. ②
Ranking
1 ②
2 ①
3 ③
Argument set:
• between March ② and May this year
• Wed 15 Sep 2021 ①
Wed 15 Sep 2021 ①
between March and May this year ②
July ③
LEAFER Module
LEAFER
LEAFER
LEAFER
Wed 15 Sep 2021
March and May
July
Layer-1: Local Understanding
Local Extractor
Local Extractor
… …
Local Extractor
Passage 1
Passage 2
Passage n

A **LEAFER** module, LEArning From ERrors, is introduced to tackle LLMs' incapability of locating exact boundaries of arguments, and yield **{a′}.**

LEAFER Module:

- A small-scale LM trained on ULTRA's errors.
- Generate insightful judgments, to rectify boundaries of candidate arguments in {a} and produce {a′}.

ULTRA: Layer-2 Self-Refinement


Pairwise Comparison & Filtering
Layer-2: Self-Refinement
Which candidate is more acceptable? A. ① B. ③
Which candidate is more acceptable? B. ② C. ③
Which candidate is more acceptable? A. ① C. ②
Ranking
1 ②
2 ①
3 ③
Argument set:
• between March ② and May this year
• Wed 15 Sep 2021 ①
Wed 15 Sep 2021 ①
between March ② and May this year
July ③
LEAFER Module
LEAFER
LEAFER
LEAFER
Wed 15 Sep 2021
March and May
July
Layer-1: Local Understanding
Local Extractor
Local Extractor
… …
Local Extractor
Passage 1
Passage 2
Passage n

Upon **{a′}**, ULTRA drops less-pertinent candidates through **self-refinement** and returns **{a^f}.**

ULTRA: Layer-2 Self-Refinement

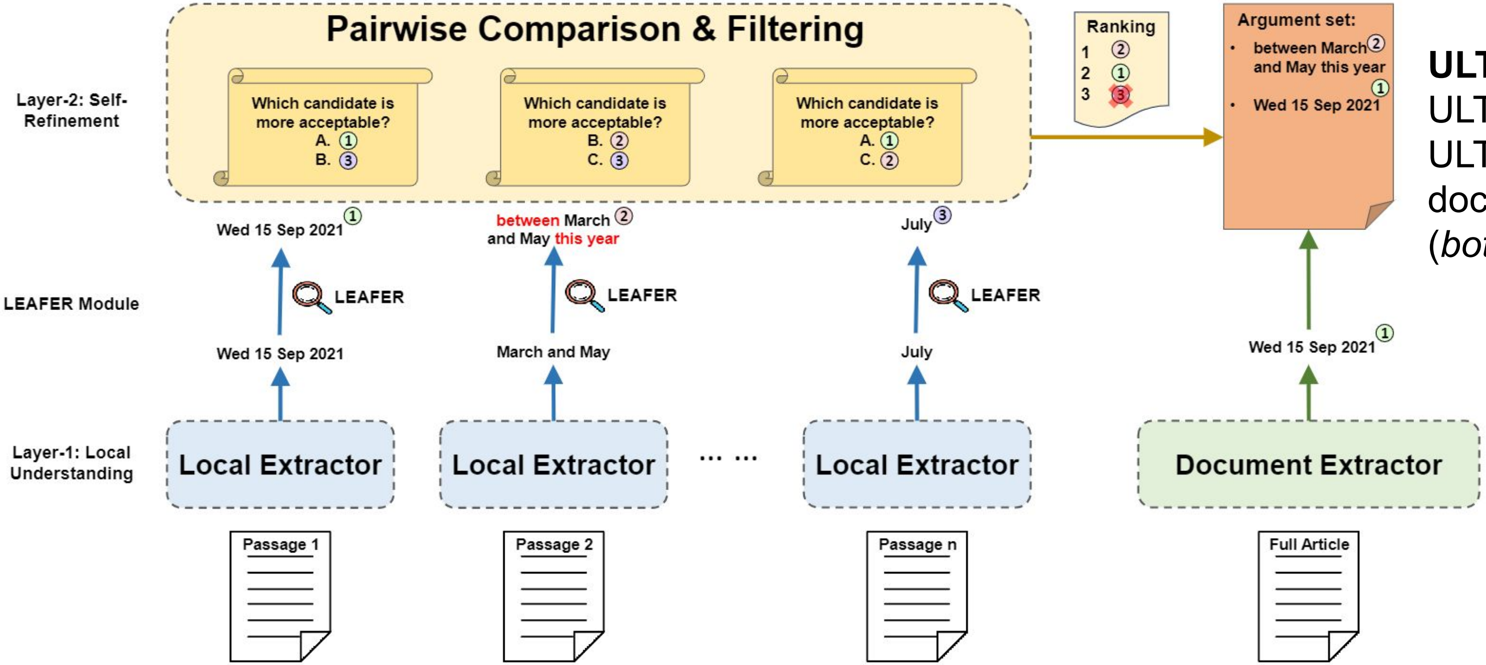
Why Self-Refinement?

- Window-based local extractors introduces **over generation** issue. To this end, we propose **ranking by pairwise comparison,** by prompting Flan-UL2 to pick a better answer between a candidate pair.

- Naïve prompting brings about two issues, and we implement solutions accordingly.

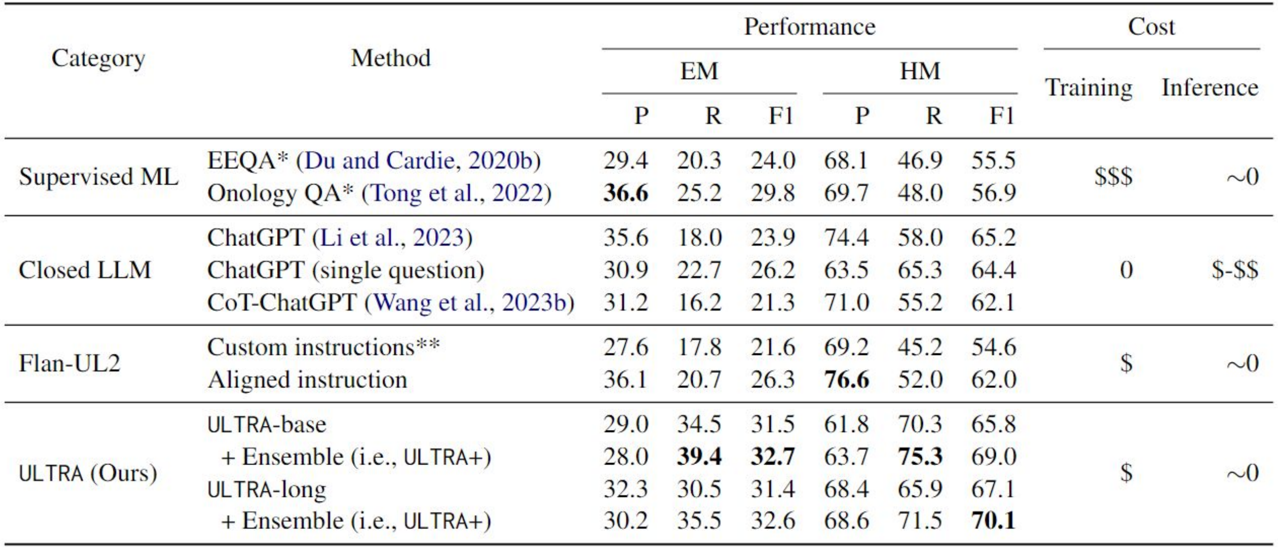
Issue	Manifestation	Solution
Positional bias	Favor candidates displayed earlier	**Calibration**: $\mathrm{P(a_i \mid d) = softmax(g(P(a_i \mid d, I; \theta), P(a_i \mid I; \theta)))}$
Lack of scalability	Quadratic growth of #comparisons	**Pruning**: “Inverted pyramid”

Model: ULTRA+



ULTRA+, a variant of ULTRA, is consisting of ULTRA (*left part*) and a document-level extractor (*bottom right*).

Results & Analysis



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**		

- Using ChatGPT for DocEAE faces two issues: **hallucination** (seemingly coherent assertions that are false in reality) and **verbosity** (extracted answers redundantly long)
- ULTRA(+) performs better across the board measured by both ***Performance*** and ***Cost***.
- ULTRA(+) also showcase the **flexibility and customizability** for accommodating various extraction criteria (*detailed in paper*).

UNIVERSITY OF MICHIGAN

Bloomberg

Engineering

Thanks!

Contact: xlfzhang@umich.edu