Document-Level Event-Argument Data Augmentation for Challenging Role Types

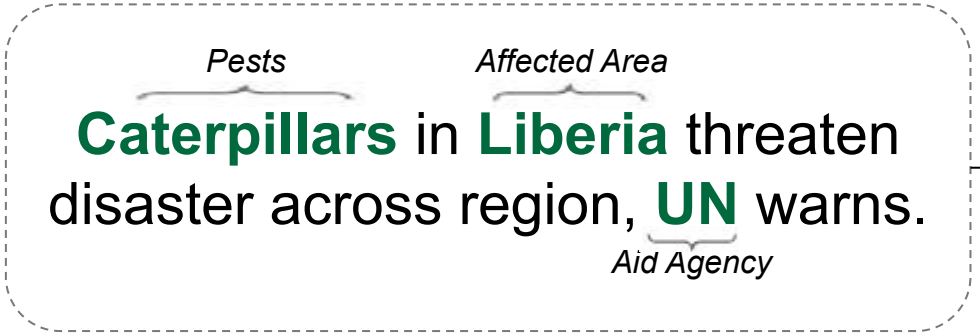
Joseph Gatto, Omar Sharif, Parker Seegmiller, Sarah M. Preum

ACL 2025

Presenter: Joseph Gatto

What Is Event Argument Extraction (EAE)?

Goal: Given an event-centric text, extract text spans corresponding to event-specific details.



Event Type: Insect Disaster

Event Roles

- **Influenced Crops and Livelihood**
- **Economic Loss**
- **Response Measures**
- **Aid Agency:** UN
- **Affected Areas:** Liberia
- **Date**
- **Pests:** Caterpillars
- **Cause**
- **Aid Supplies/Amount**

What Is Document-Level Event Argument Extraction (DocEAE)?

Goal: Extract arguments dispersed throughout **long documents**

Pests *Affected Area*

Caterpillars in **Liberia** threaten disaster across region, **UN** warns.

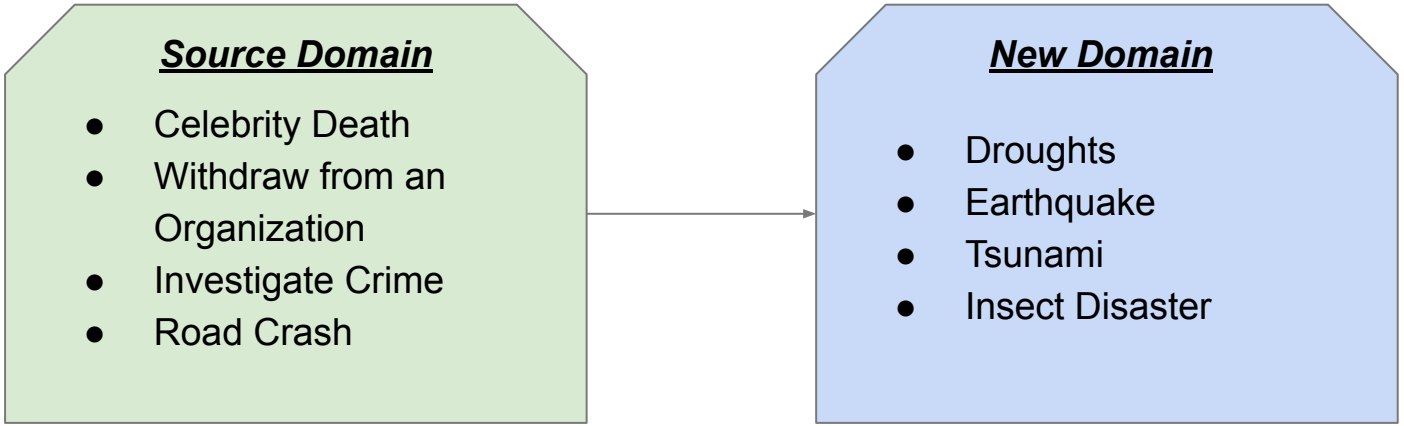
Aid Agency

Rampaging **caterpillars** in **Liberia** threaten disaster across region, warns **UN** 29 January 2009 A **United Nations** official has warned today that a UN-led team of experts is in a race against time in its attempt to halt a vast plague of caterpillars, known as armyworms, which has already swarmed across northern Liberia and threatens to march into neighbouring West African countries, destroying all crops and water supplies in its path. The enormous infestation of tens of millions of armyworms, one of the most destructive of insect pests, has forced the Liberian President, Ellen Johnson-Sirleaf, to call a national emergency in a country where access to food is already precarious. “The millions-strong caterpillar hordes devour all vegetation in their path and pollute wells and streams with their excrements wherever they go,” said Representative of the **UN** Food and Agriculture Organization (FAO) in Liberia, Winfred Hammond. Some 100 villages in northern and central Liberia have now been affected and six communities in neighbouring Guinea to the north had also been struck, in some cases overrunning buildings and sending residents fleeing in panic. According to Liberian authorities, the emergency involves about 500,000 villagers. Mr. Hammond warned that much worse could be in store as many of the caterpillars had bored into the ground, out of reach of pesticides, and formed protective cocoons around themselves, waiting to re-emerge as moths in a week or so. “Each moth can fly up to 1,000 kilometers and lay 1,000 eggs,” explained Mr. Hammond, who is an entomologist, stressing that “potentially, that's a recipe for disaster.”

Sample from DocEE Dataset

Challenges in DocEAE

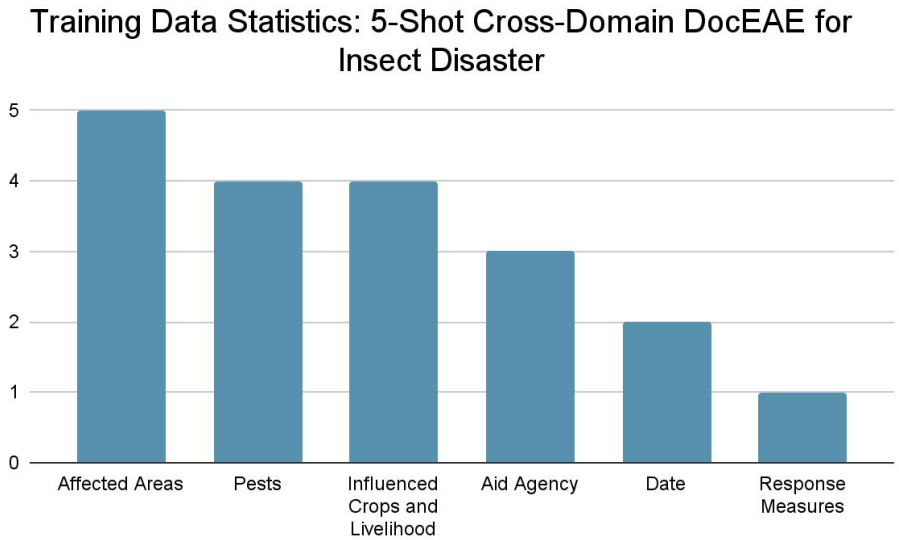
Challenge 1: Obtaining data for new event types is extremely difficult and expensive.



Need Methods for Improved Low-Resource, Cross-Domain DocEAE!

Challenges in DocEAE

Challenge 2: Difficult to model rare role types



Event Type: Insect Disaster

- **Influenced Crops and Livelihood**
- **Economic Loss: MISSING**
- **Response Measures**
- **Aid Agency**
- **Affected Areas**
- **Date**
- **Pests**
- **Cause: MISSING**
- **Aid Supplies/Amount: MISSING**

Solution: Data Augmentation for DocEAE

Past Studies Largely Focus on

1. Augmentation for Sentence-Level EAE
2. Augmenting existing samples
3. Applying pre-trained classifiers to unlabeled corpora

Our Goal

1. Make LLMs useful for generating DocEAE samples.
2. Remove dependence on in-domain data
3. Improve representation of new events and rare role types

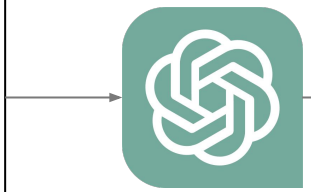
LLMs For DocEAE Data Generation

Challenges

Why Don't LLMs Work Off-The-Shelf for DocEAE Data Generation?

Event Type: Insect Disaster

- **Influenced Crops and Livelihood**
- **Economic Loss**
- **Response Measures**
- **Aid Agency:** UN
- **Affected Areas:** Liberia
- **Date**
- **Pests:** Caterpillars
- **Cause**
- **Aid Supplies/Amount**



Pests *Affected Area*

Caterpillars in **Liberia** threaten disaster across region, **UN** warns.

Aid Agency

LLMs For DocEAE Data Generation

Challenges

Why Don't LLMs Work Off-The-Shelf for DocEAE Data Generation?

Issue #1: Under Labeling

LLMs are highly likely to generate role information which was not requested!

Prompt

Generate Document with Following Event Structure:
{Pests: Caterpillars, Affected Area: Liberia}

Pests *Affected Area*

Caterpillars in **Liberia** threaten disaster across region, **UN** warns.

LLMs For DocEAE Data Generation

Challenges

Why Don't LLMs Work Off-The-Shelf for DocEAE Data Generation?

Issue #2: Trouble Generating Substrings

We need to generate a document suitable for an **extractive task!** LLMs struggle to follow such instructions.

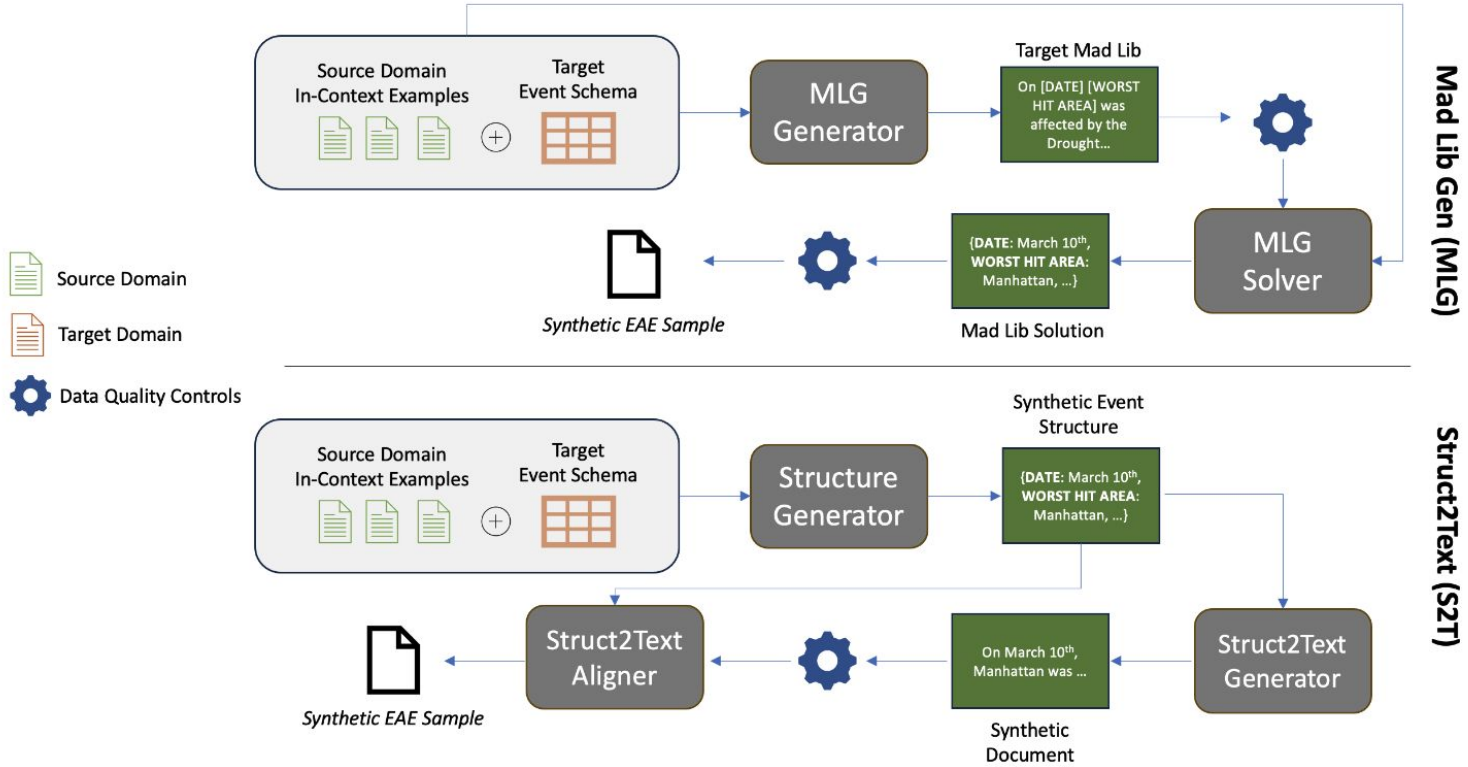
Prompt

Generate Document with Following Event Structure:
{Pests: Caterpillars, Affected Area: Liberia, Aid Agency: United Nations}

Pests *Affected Area*

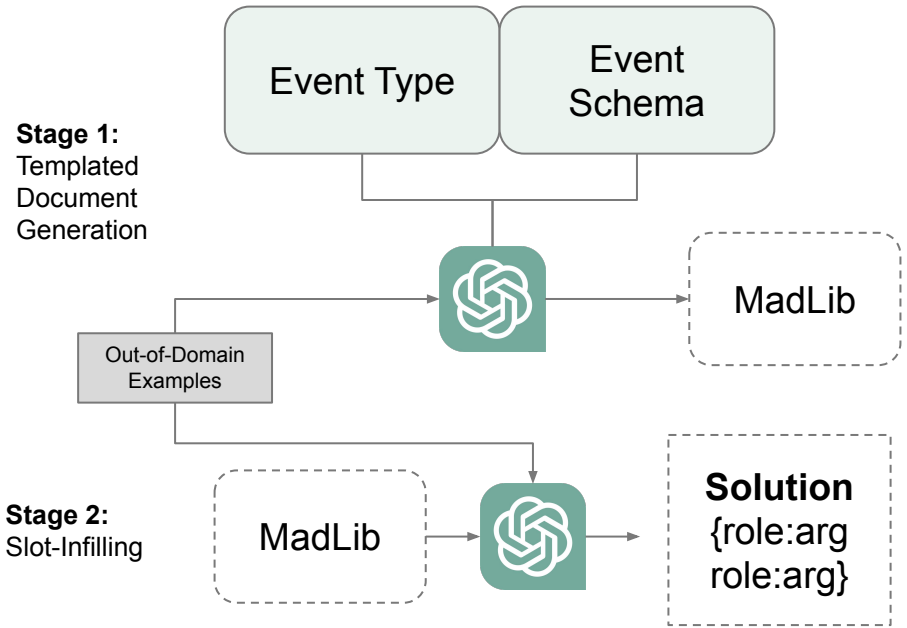
Caterpillars in **Liberia** threaten disaster across region, **UN** warns.

Generating Synthetic Data



LLMs For DocEAE Data Generation

Solution #1: Templated Document Generation (Mad Lib Gen)



MadLib

[Pests] in [Affected Area]
threaten disaster across region
[Aid Agency] warns

Solution

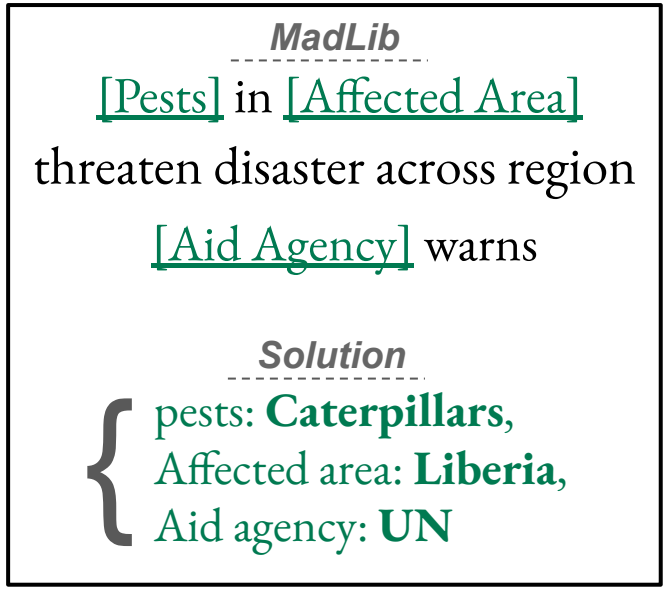
{ pests: **Caterpillars**,
Affected area: **Liberia**,
Aid agency: **UN**

Mad Lib Generation and Solving

LLMs For DocEAE Data Generation

Solution #1: Templated Document Generation (Mad Lib Gen)

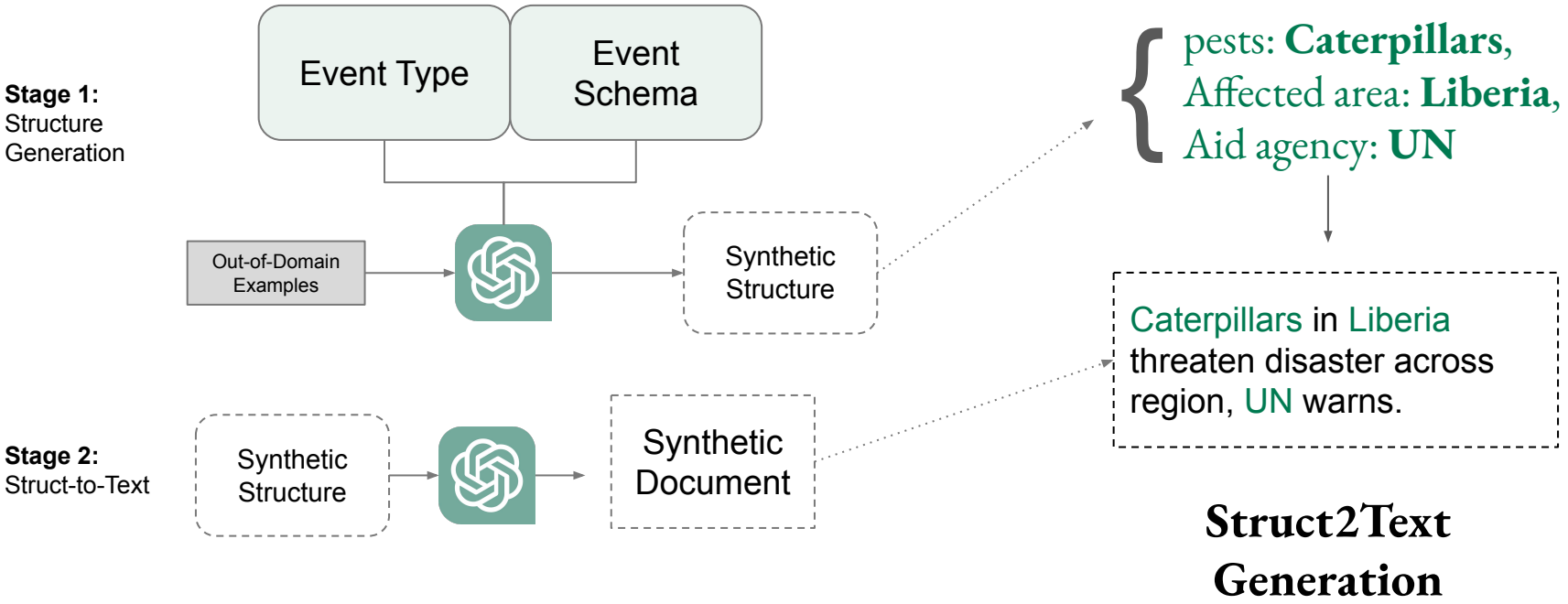
- We solve **Under Labeling** by providing the full event schema during generation
- We solve **generation for span extraction** by created templated documents which can be string replaced for document construction



Mad Lib Generation and Solving

LLMs For DocEAE Data Generation

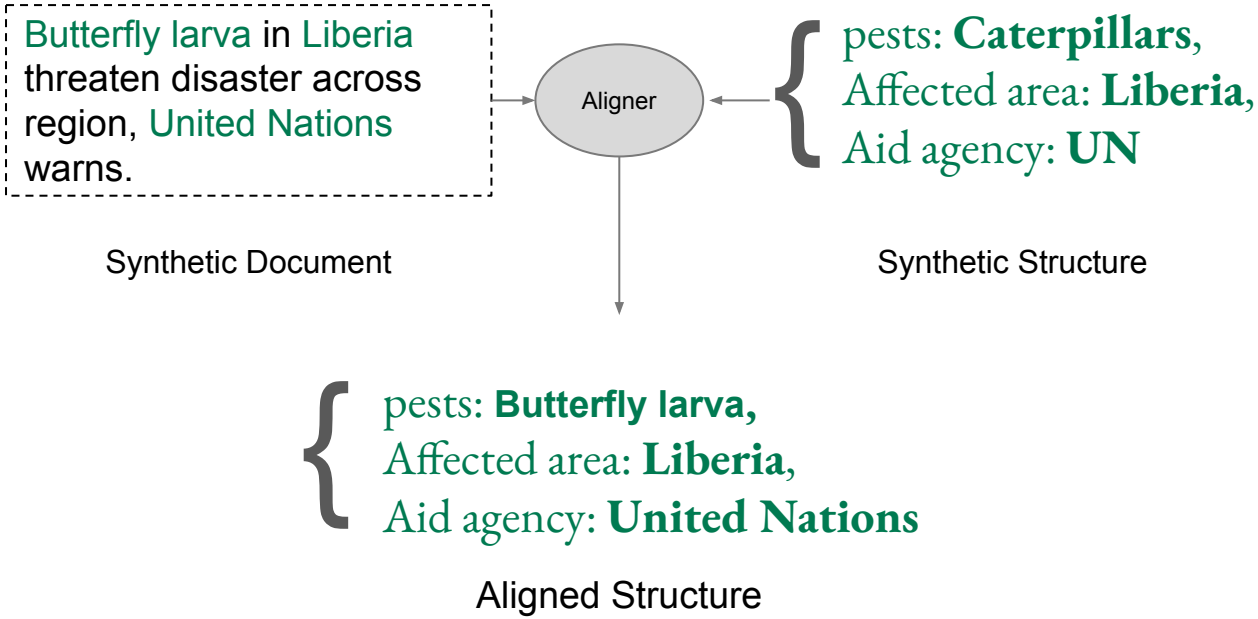
Solution #2: Struct-to-Text Generation with Span Realignment



LLMs For DocEAE Data Generation

Solution #2: Struct-to-Text Generation with Span Realignment

Stage 3:
Text-Structure Alignment



Experiments

Task: Few-Shot, Cross-Domain DocEAE

Dataset: DocEE[1]

Example Source Domain Events:

- Archeological Discoveries
- Famous Person - Give a Speech
- Organization Merge

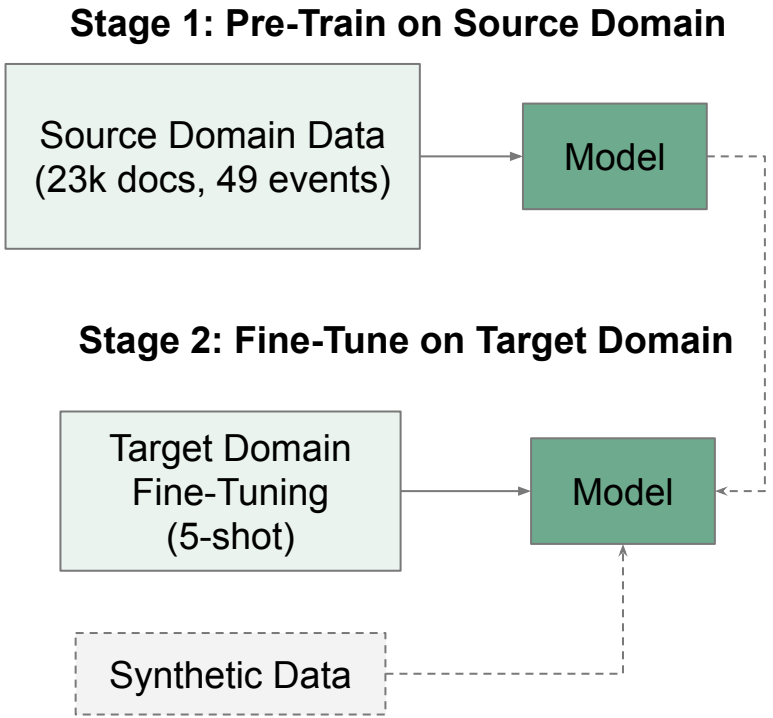
Target Domain Theme: Natural Disasters (E.g. Earthquakes, Tsunamis)

Split	# Samples	# Events
Source Train	23,630	49
Target Train	50	10
Target Dev	1,850	10
Target Test	1,955	10

DocEE Dataset Statistics

1. DocEE: A Large-Scale and Fine-grained Benchmark for Document-level Event Extraction (Tong et al., NAACL 2022)

Experiments



Evaluation Metrics

Model Output = (**Document ID**, **Role**, **Argument**)

Metric	Description
F1	Standard F1 On All Data
Role F1	Compute F1 For *Each Role Individually*, Then Take Average
0-Shot F1	Role F1 for 0-Shot Roles (n=9)
Role-Depth F1 (RDF1)	Role F1 for **Semantically Outlying Roles** (n=16)

Baselines

Baselines Models

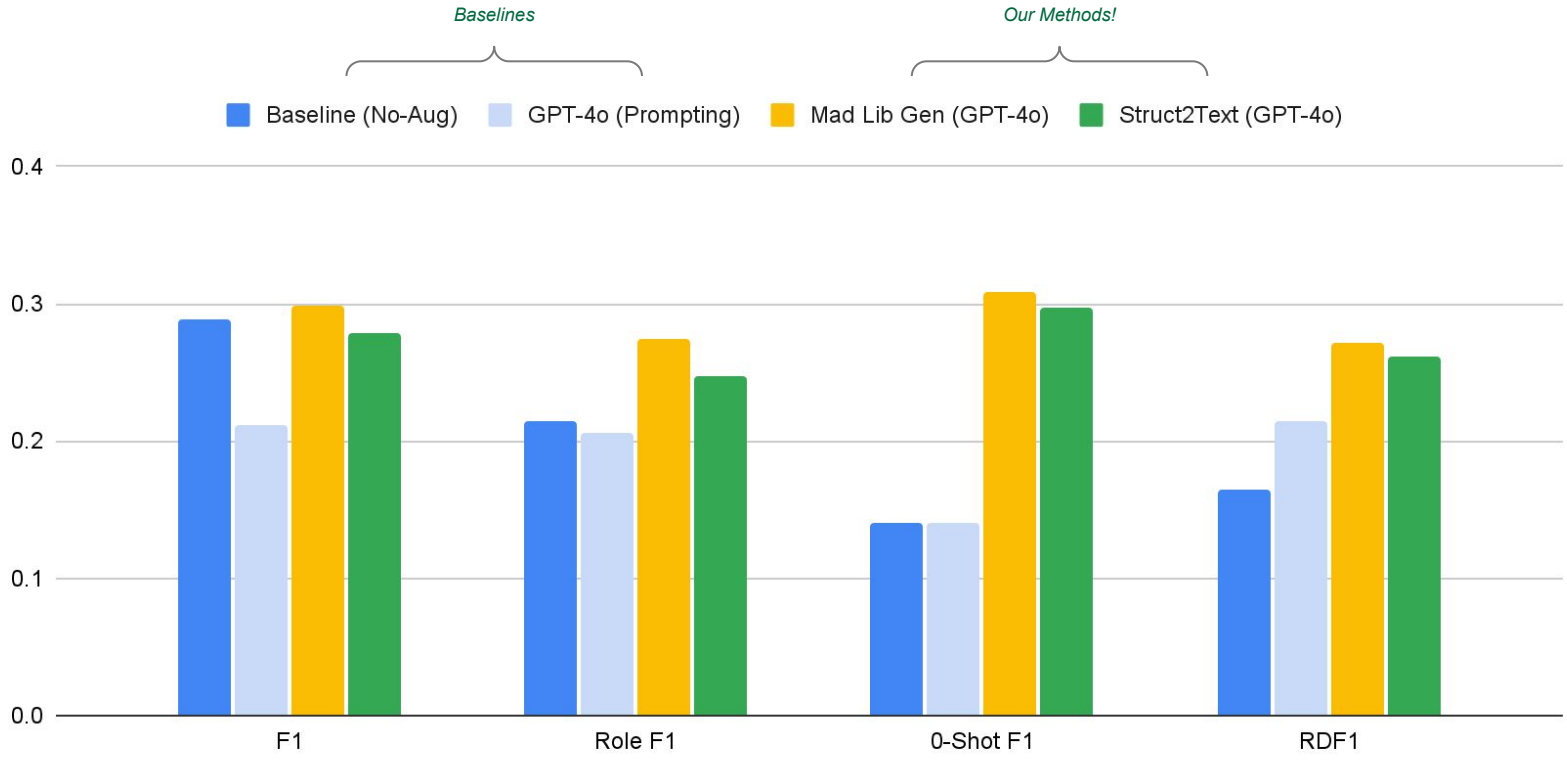
- **BERT-QA:** BERT-based Span Extraction
- **GPT-4o:** LLM Prompting for EAE

Our Methods

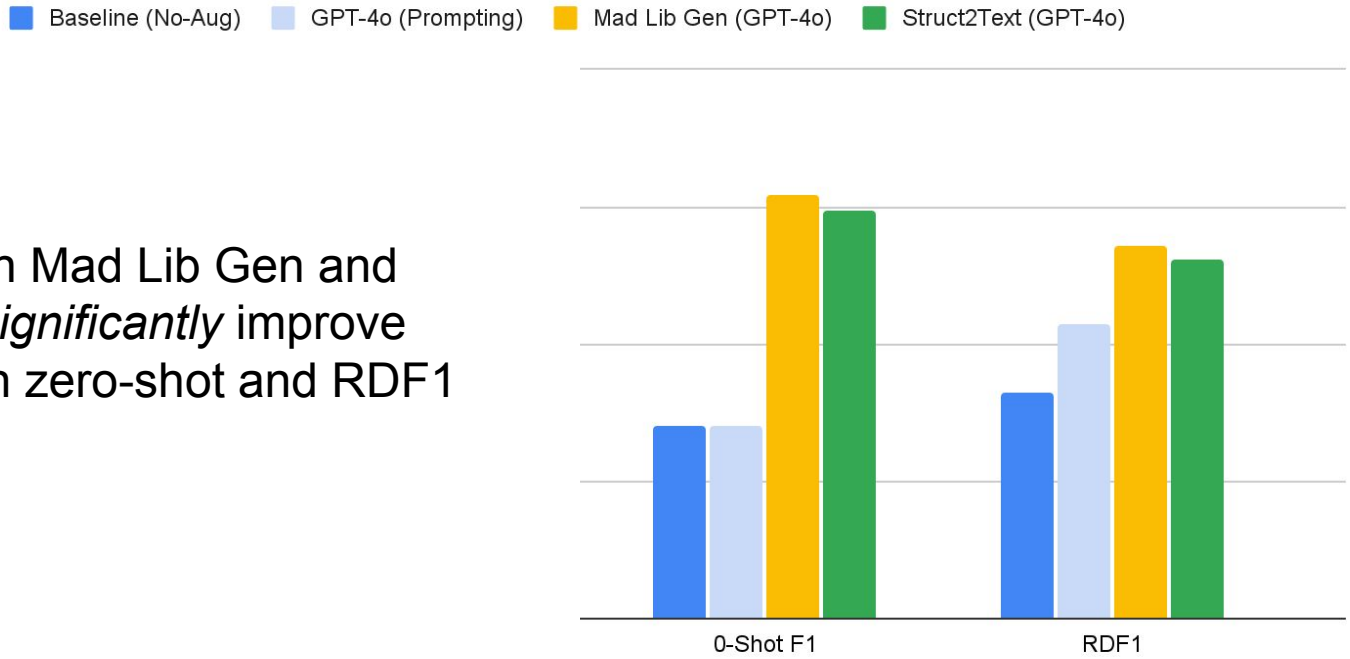
- **Mad Lib Gen (MLG):** GPT-4o backbone
- **Struct2Text (S2T):** GPT-4o backbone

**Please see our paper for results on additional models!*

Results: BERT-QA

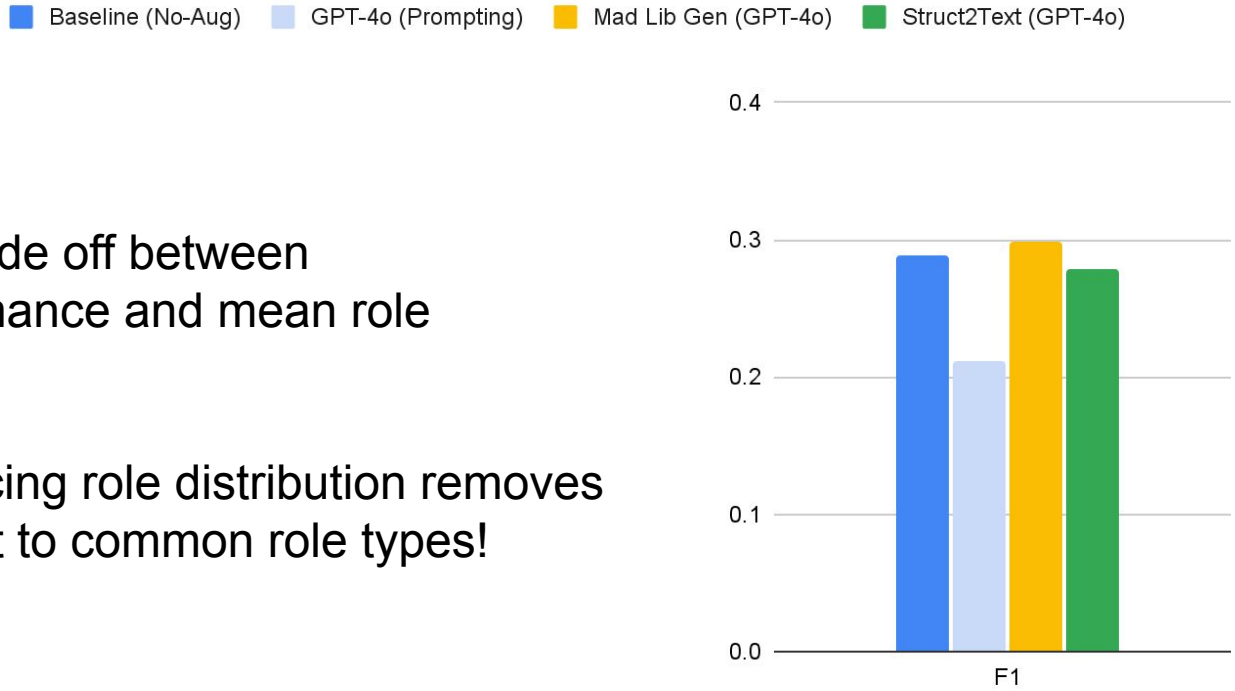


Results: BERT-QA



Result #1: Both Mad Lib Gen and Struct-to-Text *significantly* improve performance on zero-shot and RDF1 roles.

Results: BERT-QA



Result #2: Trade off between overall-performance and mean role performance

Cause: Balancing role distribution removes ability to overfit to common role types!

Results: MLG Boosts Performance on Other EAE Datasets!

Result #3: MLG augmentation boosts performance in low-resource settings on other tasks!

	F1	Role F1
DiscourseEE (10%)	0.13	0.059
DiscourseEE (10%) + Aug	0.334†	0.342†
DiscourseEE (50%)	0.163	0.092
DiscourseEE (50%) + Aug	0.361†	0.357†
DiscourseEE (Full)	0.18	0.106
DiscourseEE (Full) + Aug	0.403†	0.396†
RAMS (10%)	0.134	0.128
RAMS (10%) + Aug	0.214†	0.186†
RAMS (50%)	0.323	0.298
RAMS (50%) + Aug	0.343†	0.33†
RAMS (Full)	0.388	0.38
RAMS (Full) + Aug	0.393	0.375
PHEE (10%)	0.42	0.303
PHEE (10%) + Aug	0.544†	0.53†
PHEE (50%)	0.596	0.581
PHEE (50%) + Aug	0.599	0.594
PHEE (Full)	0.621	0.618
PHEE (Full) + Aug	0.618	0.608

Table: BERT-QA results showing performance with and without 500 synthetic samples + x% of training data for each of DiscourseEE, RAMS, and PHEE.

Conclusion: MLG vs S2T → Which to use?

Mad Lib Gen (MLG)

Benefits

★ More open-source friendly. Works with smaller LLMs!

★ More accurate span annotations.

Limitations

- Difficult to include additional event information.

Struct2Text (S2T)

Benefits

★ Straightforward to include additional event information!

Limitations

- Relies on Langchain + OpenAI Pydantic integration. Not small-LLM friendly.

Thank You!

Document-Level Event-Argument Data Augmentation for Challenging Role Types

Joseph Gatto, Omar Sharif, Parker Seegmiller, Sarah M. Preum