Should We Transform Coreference Resolution into a Text Generation Task?

CRAC 2025 Shared Task Submissions

Antoine Bourgois, Olga Seminck, Yoann Dupont, Mathieu Dehouck, Marine Delaborde

Lattice Laboratory (CNRS – École Normale Supérieure), Paris, France

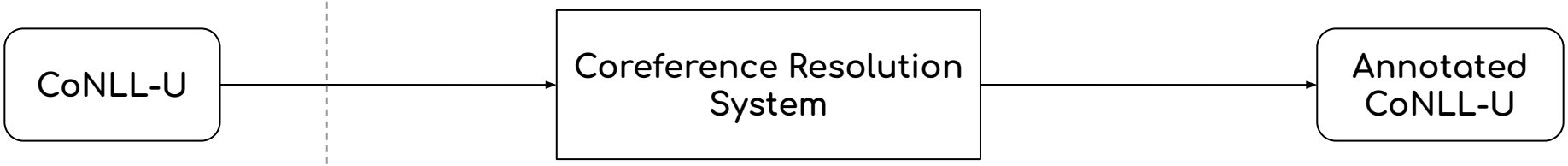
Eighth Workshop on Computational Models of Reference, Anaphora and Coreference - CRAC 2025

Shared Task Ressources

- CorefUD 1.3 Corpus
- 17 Languages
  Ancient Greek, Biblical Hebrew, Catalan, Czech, English, French, German, Hindi, Hungarian, Korean, Lithuanian, Norwegian, Old Church Slavonic, Polish, Russian, Spanish, Turkish
- 22 Datasets
  - ↳ Documents
    - ↳ Sentences

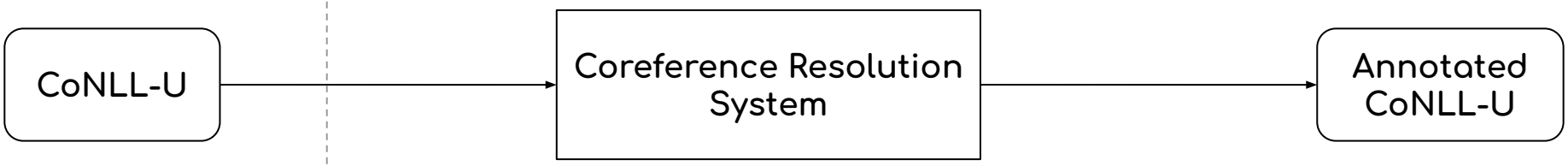
Shared Task Goal

22 Datasets



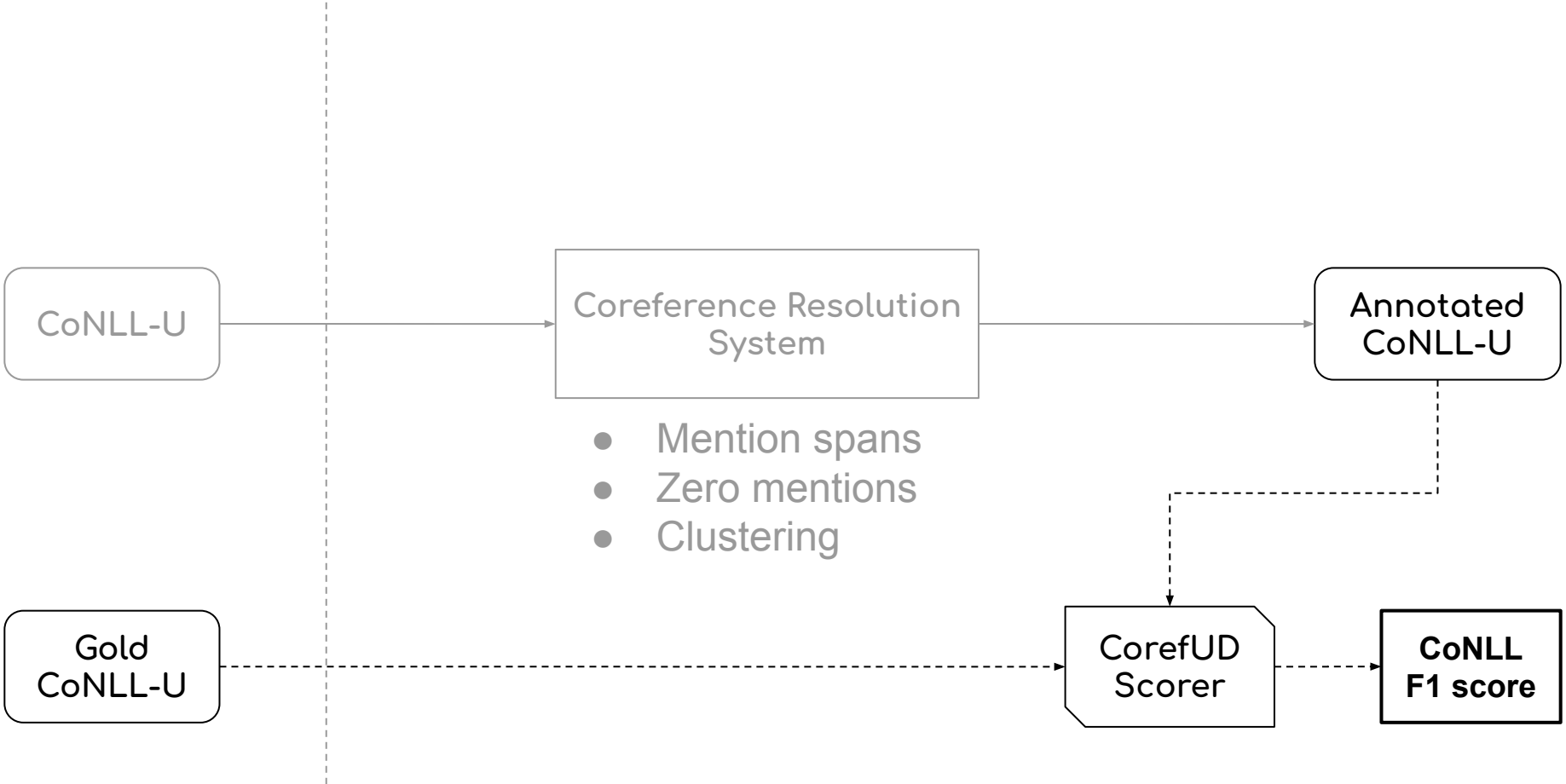
Shared Task Goal

22 Datasets

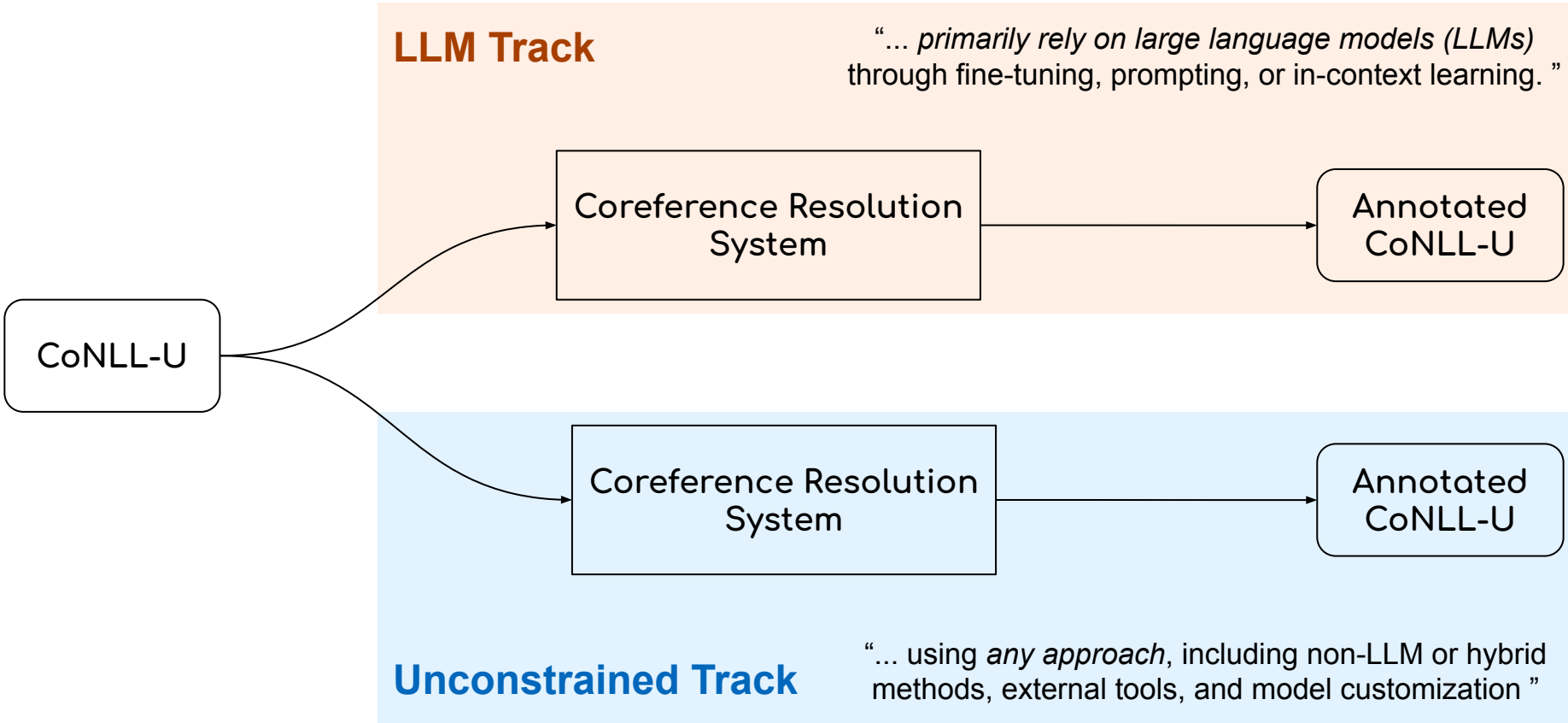


- Mention spans
- Zero mentions
- Clustering

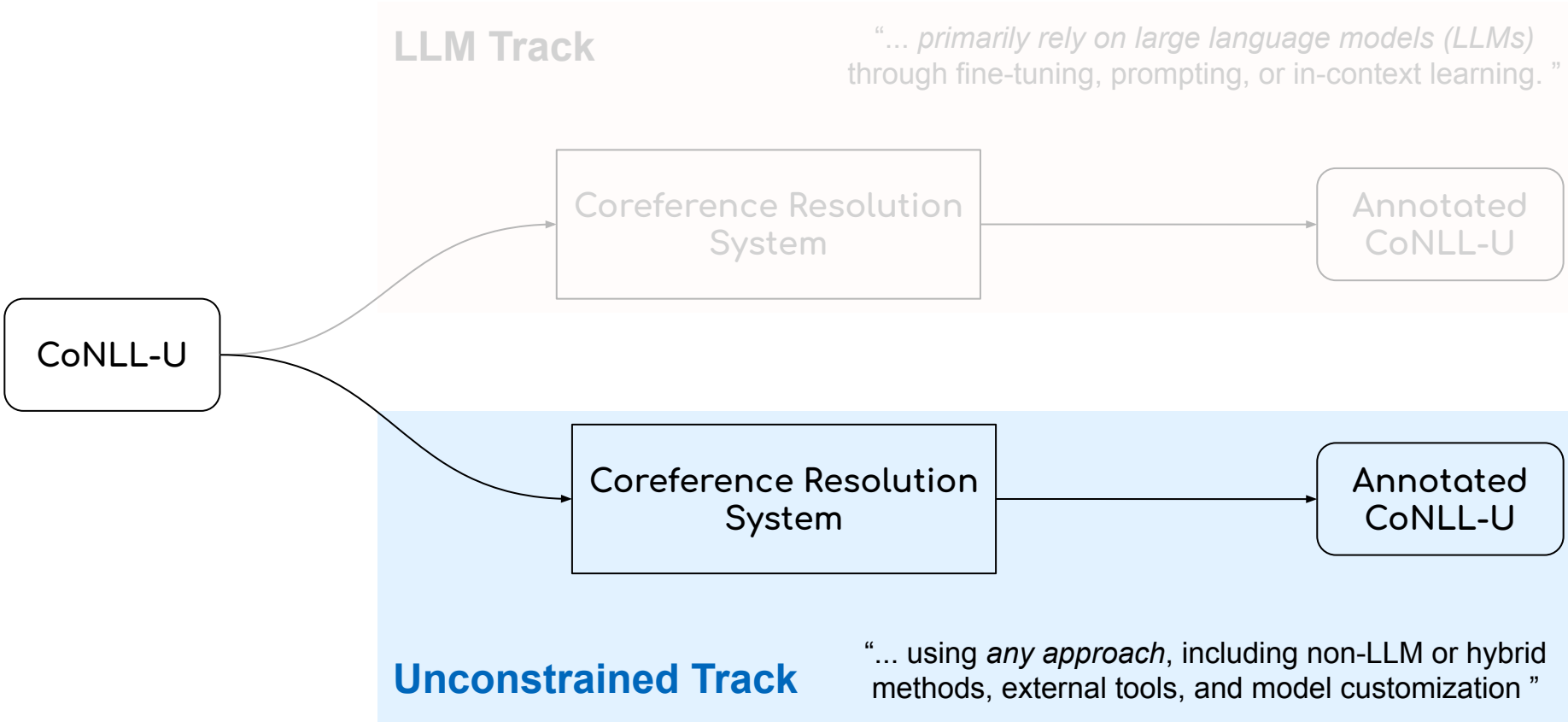
Shared Task Evaluation



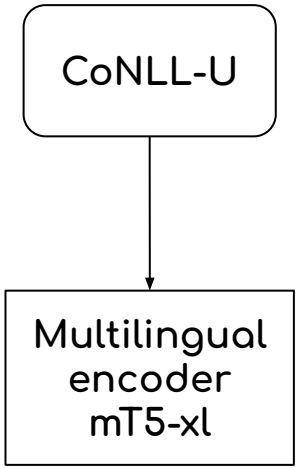
LLM & Unconstrained Tracks



Unconstrained Track



Unconstrained Submission: Multistage Pipeline



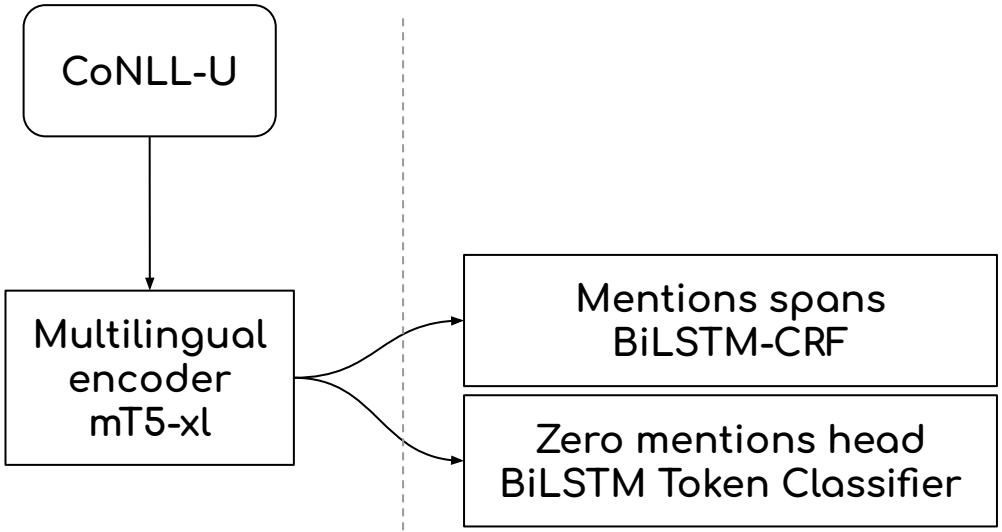
Token Embeddings

Mentions Detection

Mention Pairs

Clustering

Unconstrained Submission: Multistage Pipeline



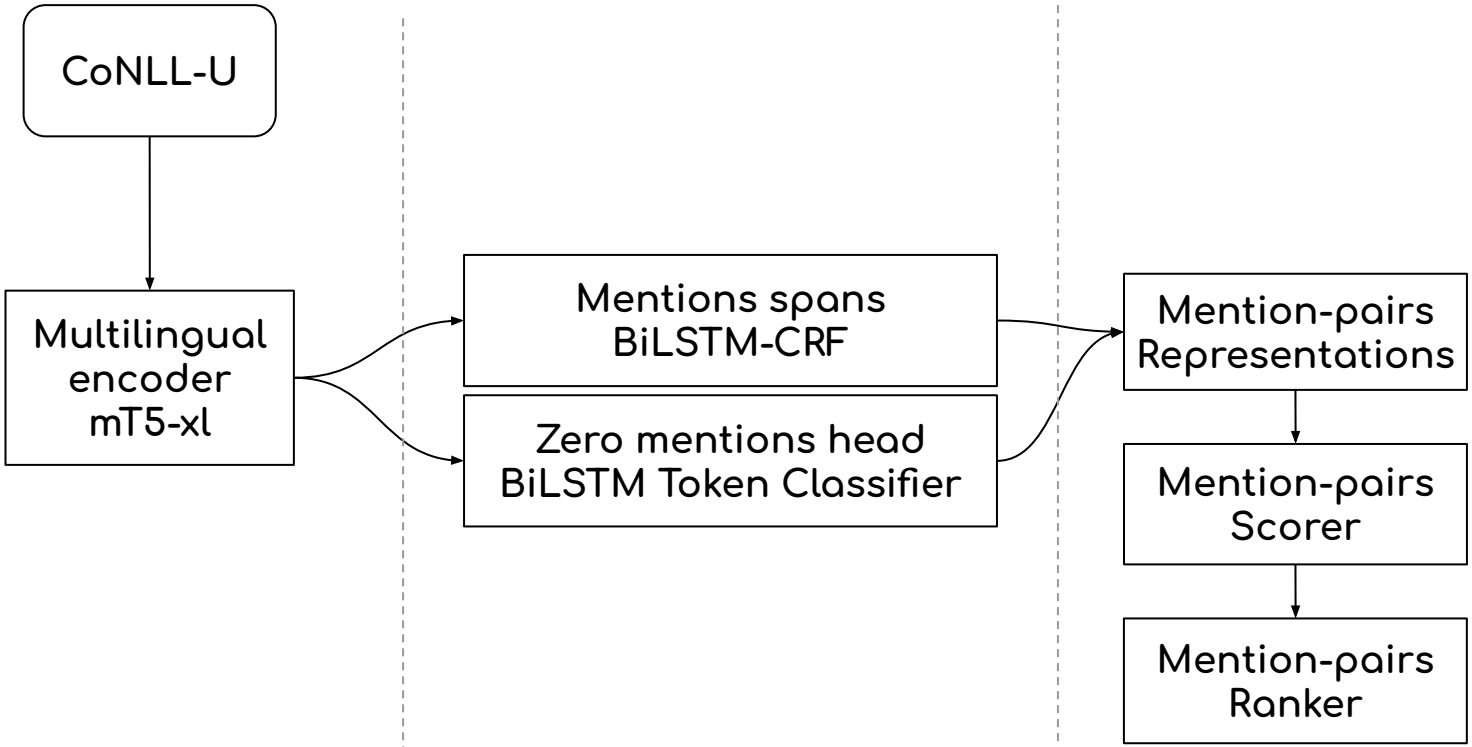
Token Embeddings

Mentions Detection

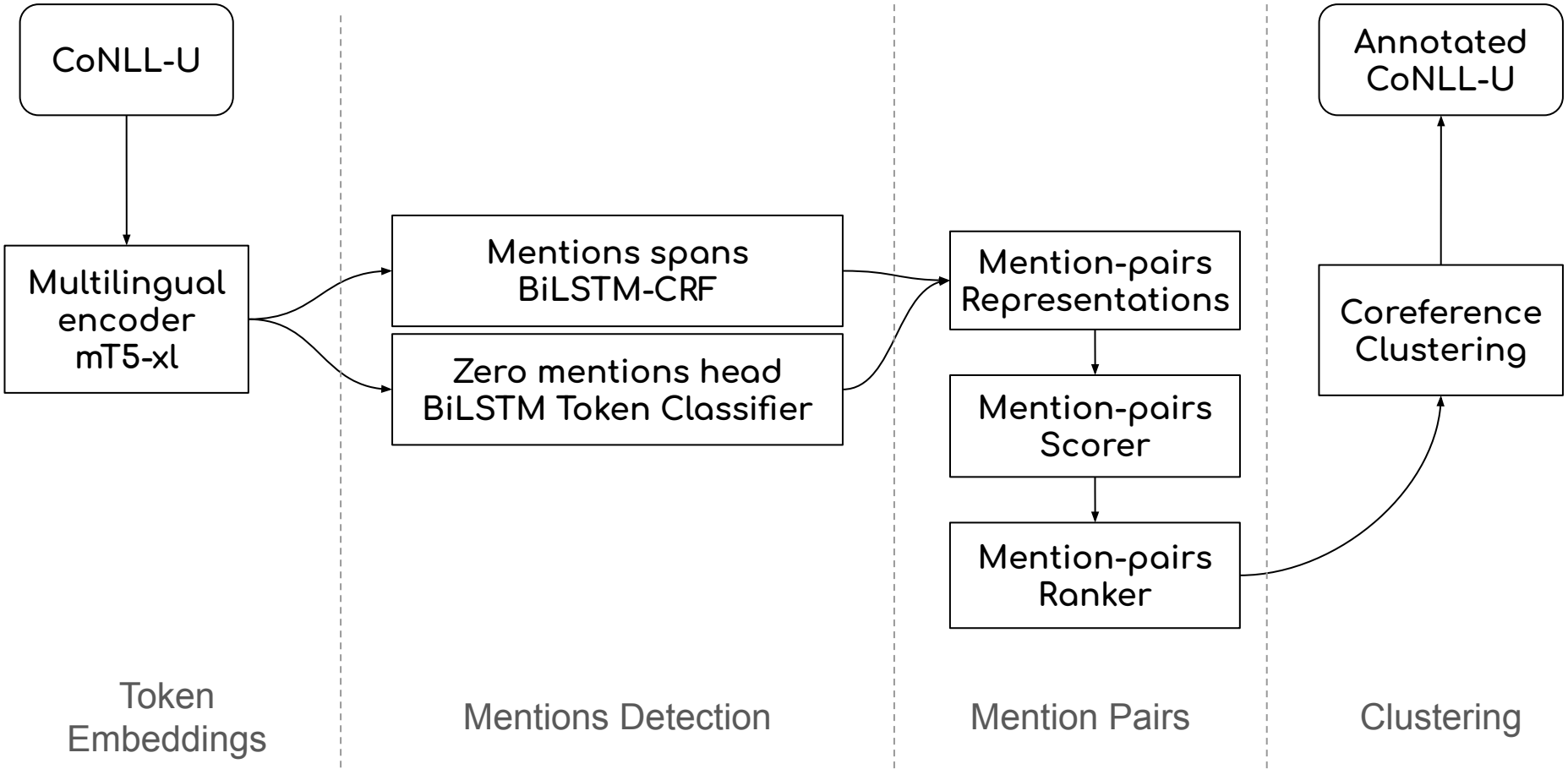
Mention Pairs

Clustering

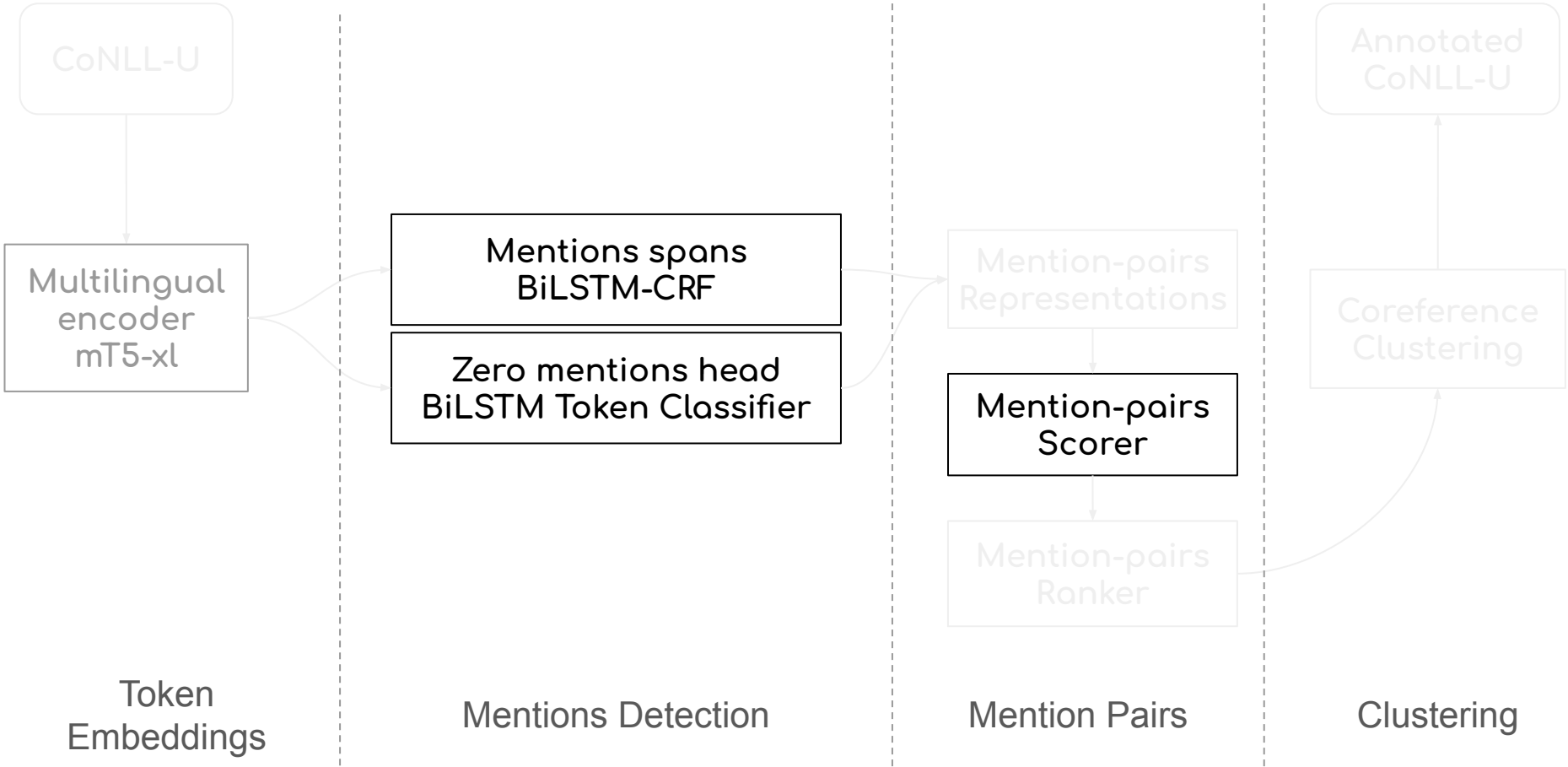
Unconstrained Submission: Multistage Pipeline



Unconstrained Submission: Multistage Pipeline



Unconstrained Submission: Trained Modules

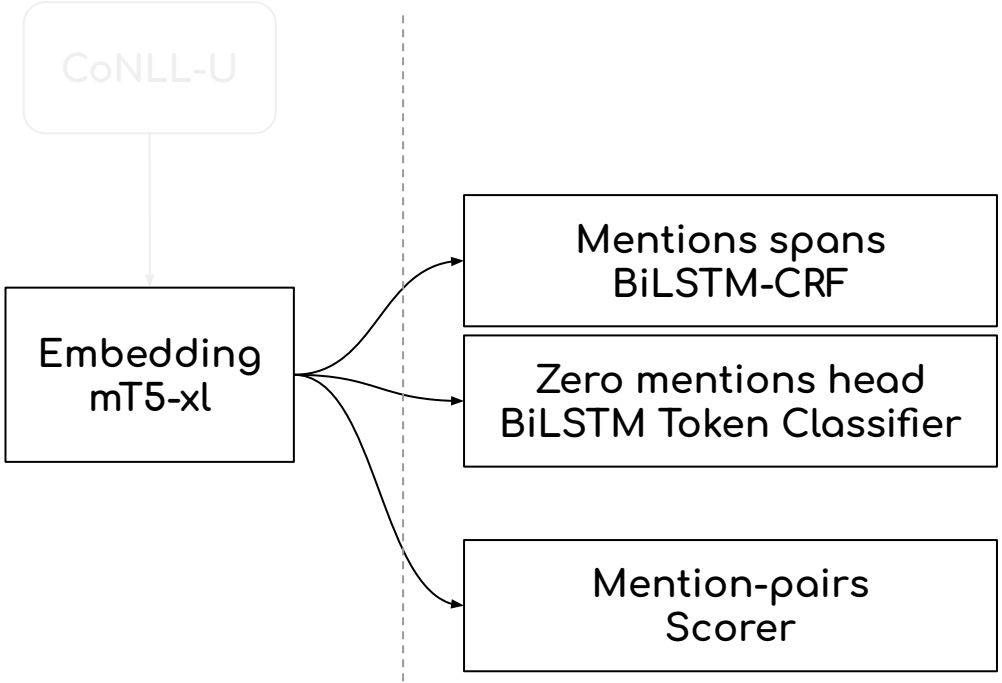


Unconstrained Submission: Training Resources

Token Embeddings

	GPU Peak VRAM	Training Time
Mentions spans BiLSTM-CRF	3.8 GiB	< 5 hours
Zero mentions head BiLSTM Token Classifier	1.7 GiB	< 4 hours
Mention-pairs Scorer	1.7 GiB	< 3 hours
	7.2 GiB	**< 5 hours**

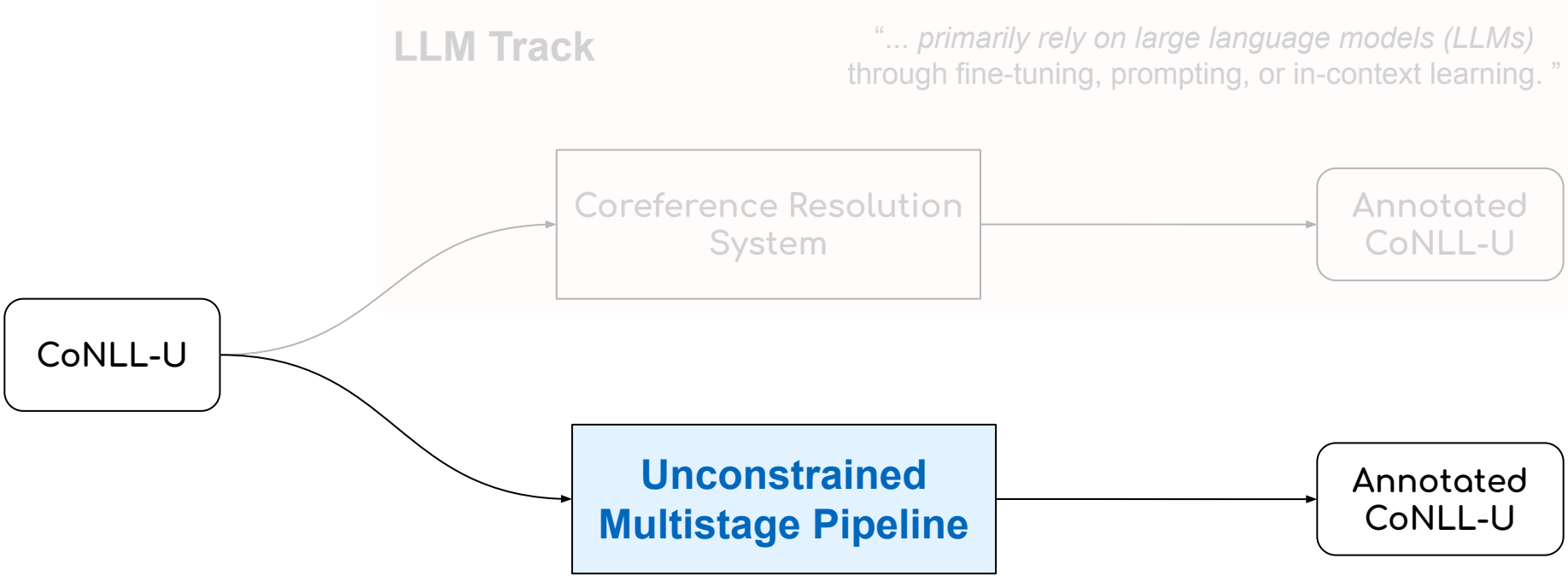
Unconstrained Submission: Training Resources



	GPU Peak VRAM	Training Time
Mentions spans BiLSTM-CRF	3.8 GiB	< 5 hours
Zero mentions head BiLSTM Token Classifier	1.7 GiB	< 4 hours
Mention-pairs Scorer	1.7 GiB	< 3 hours
	7.2 GiB	**< 5 hours**

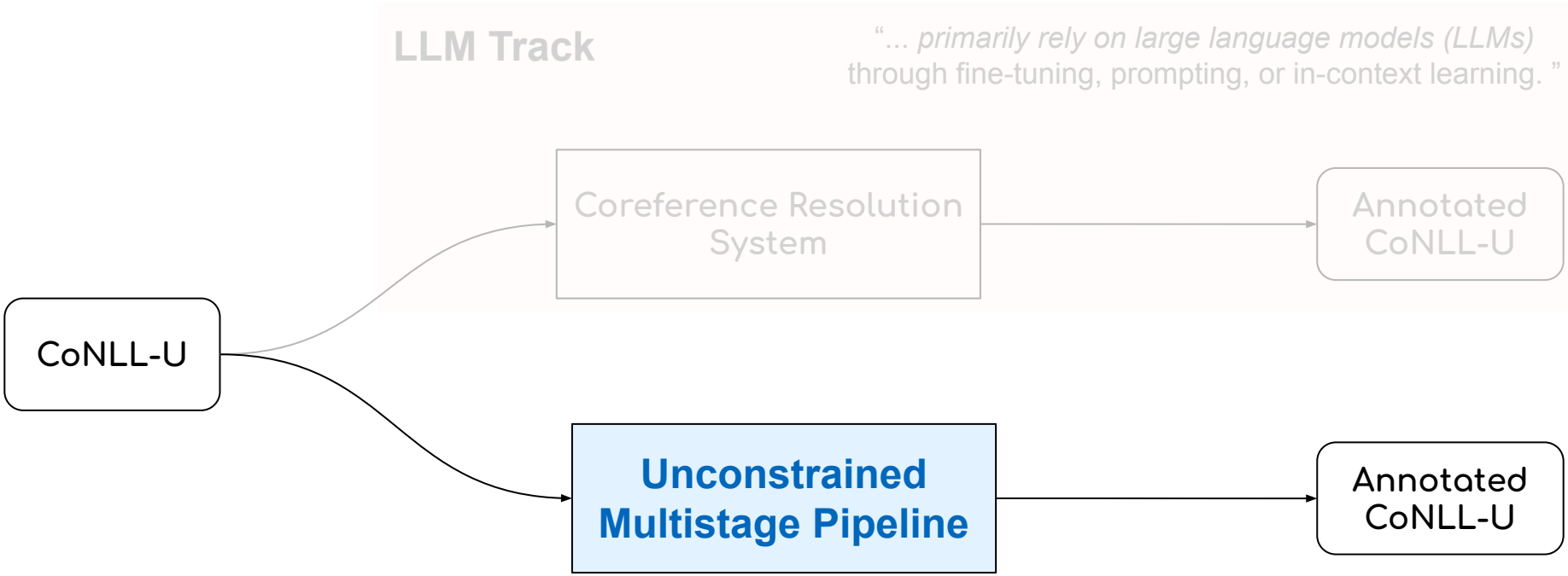
- **8 GiB**
- **< 1 hour**

Unconstrained Submission: Resources



Train	8 GiB	6 hours
Dev	8 GiB	16 minutes

Unconstrained Submission: Performance

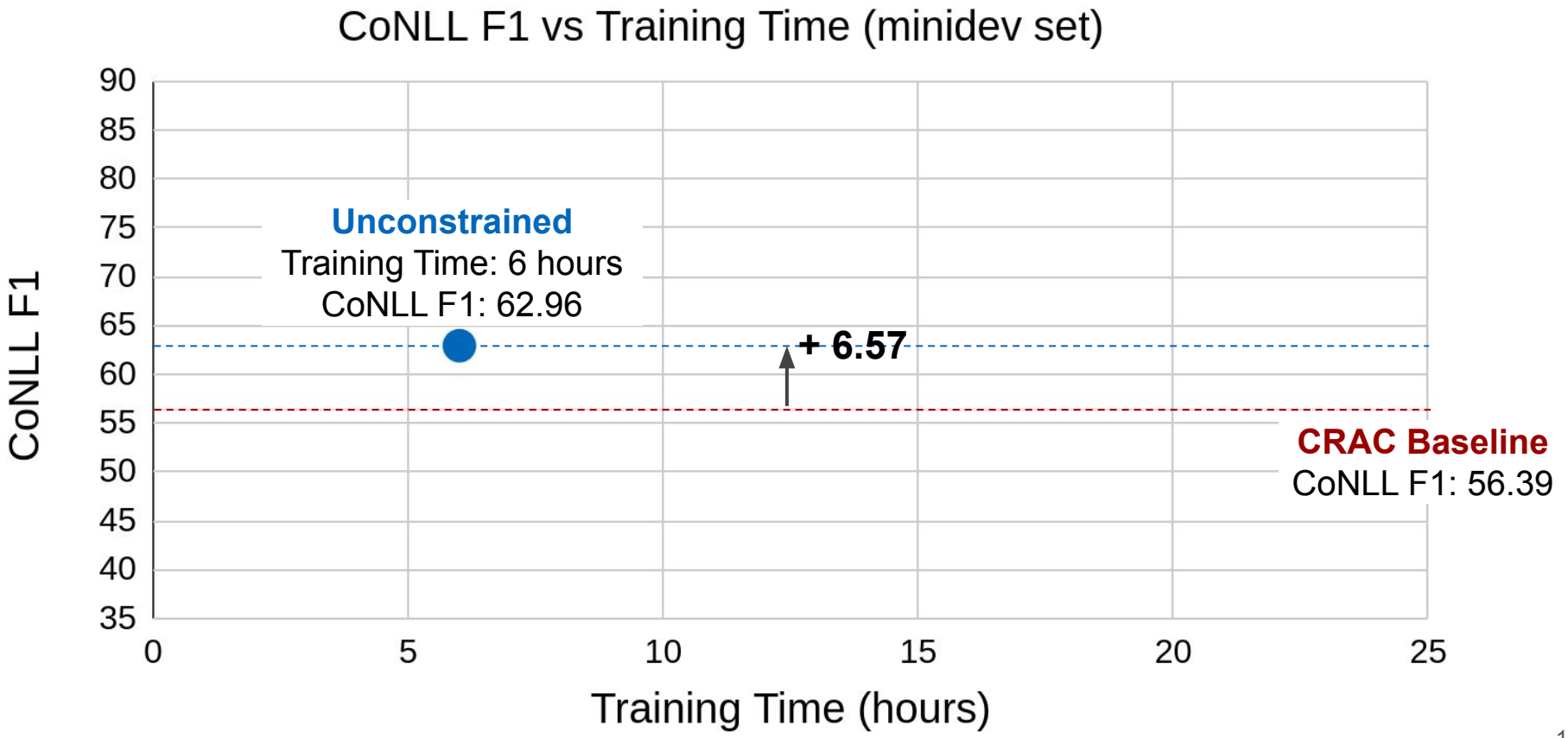


Train	8 GiB	6 hours
Dev	8 GiB	16 minutes

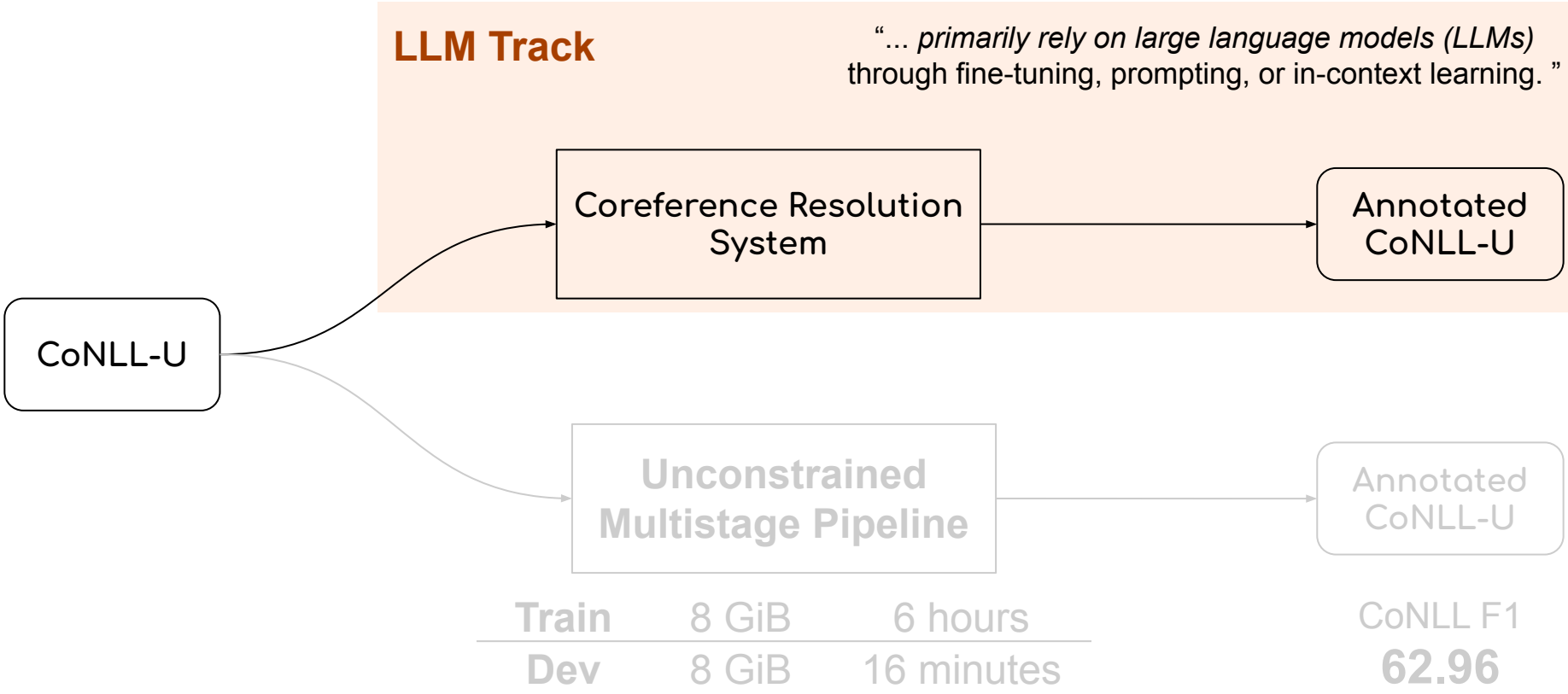
CoNLL F1

≈ 62.96

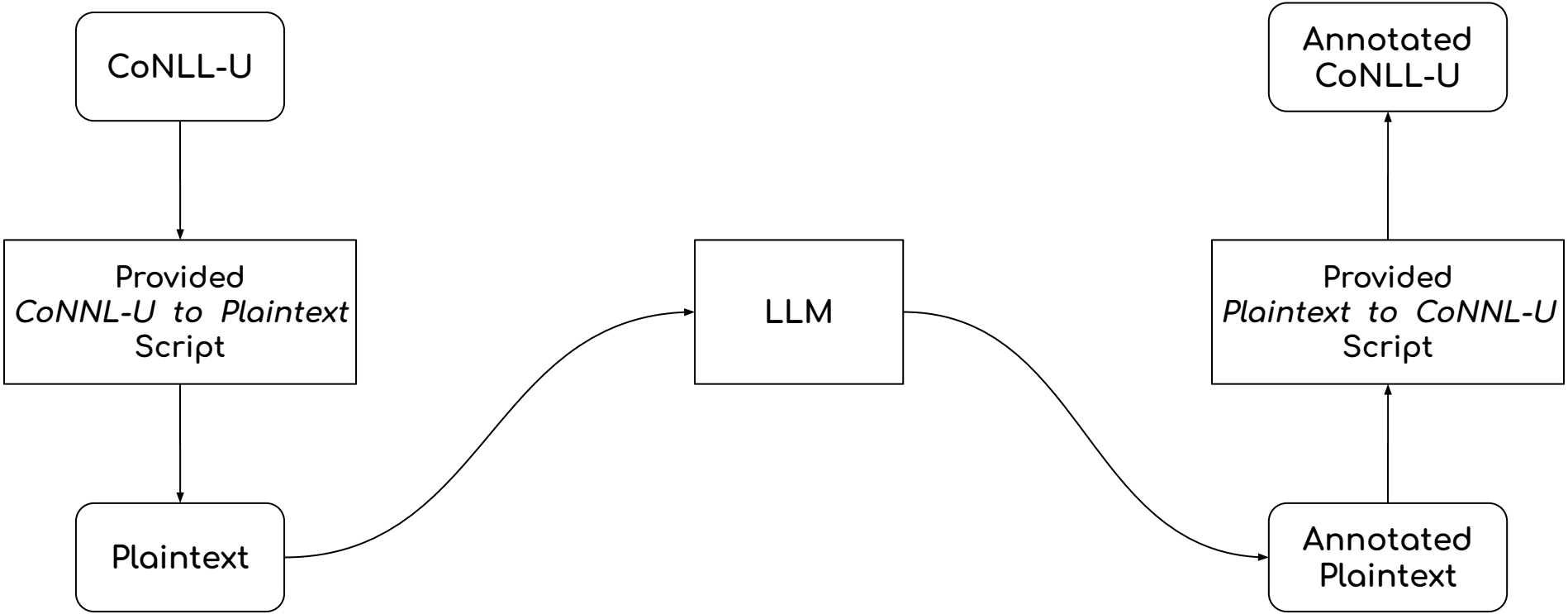
Comparison of Models



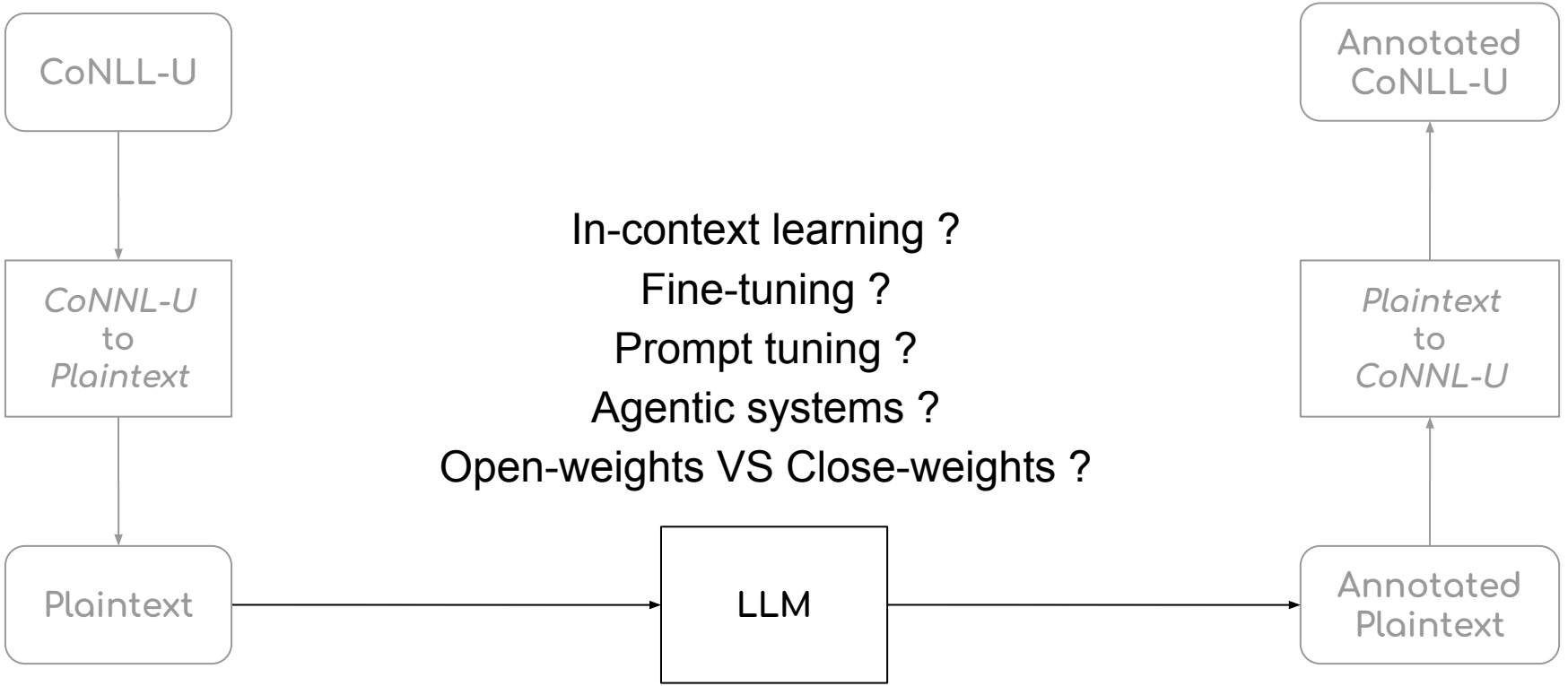
LLM Submission



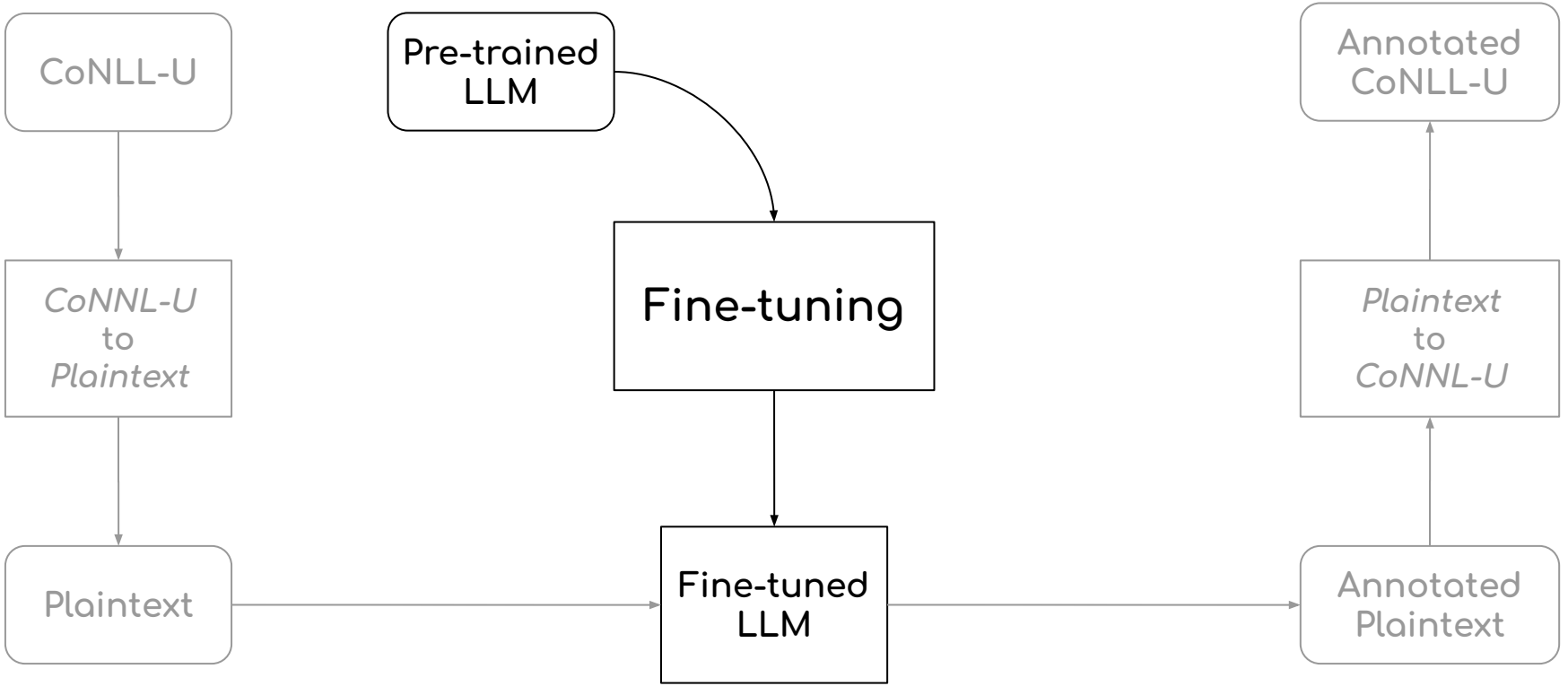
LLM Submission: Provided Conversion Scripts



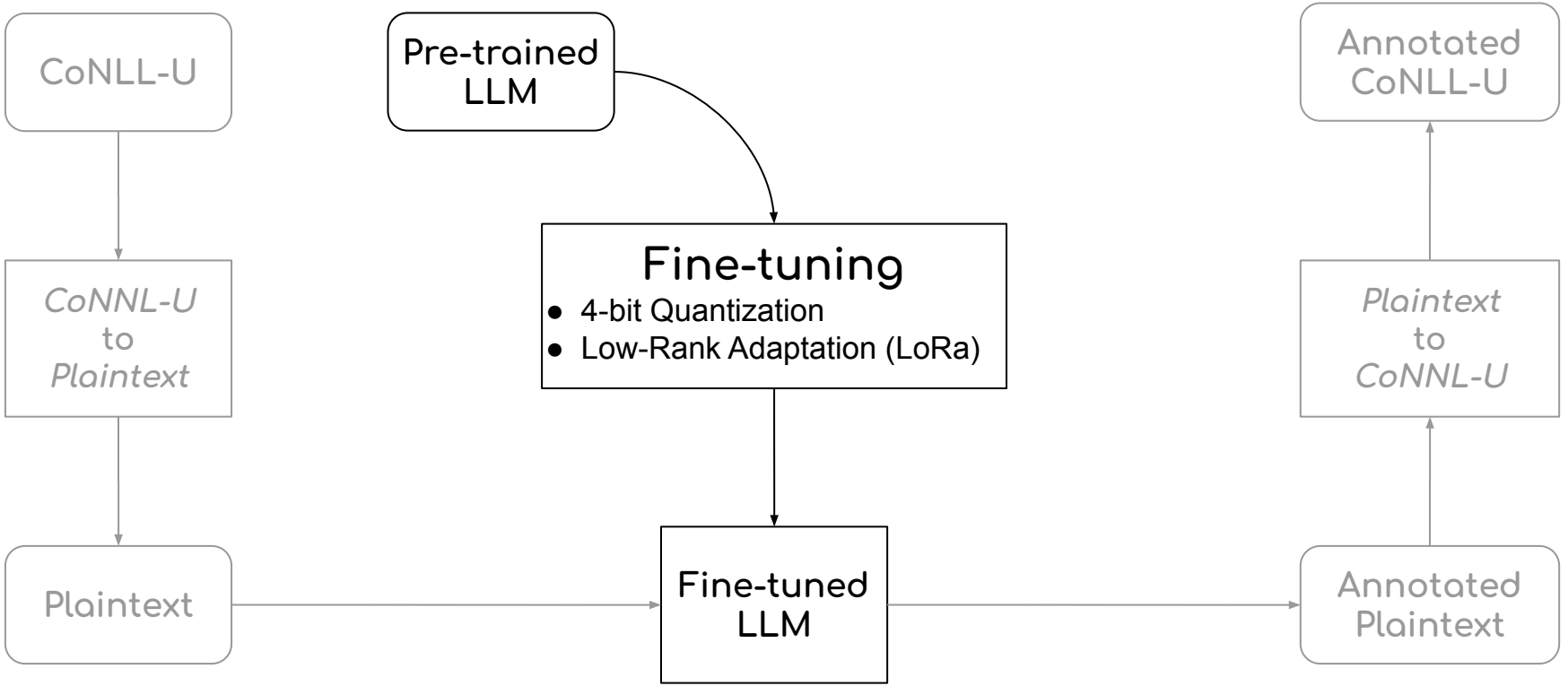
LLM Strategies



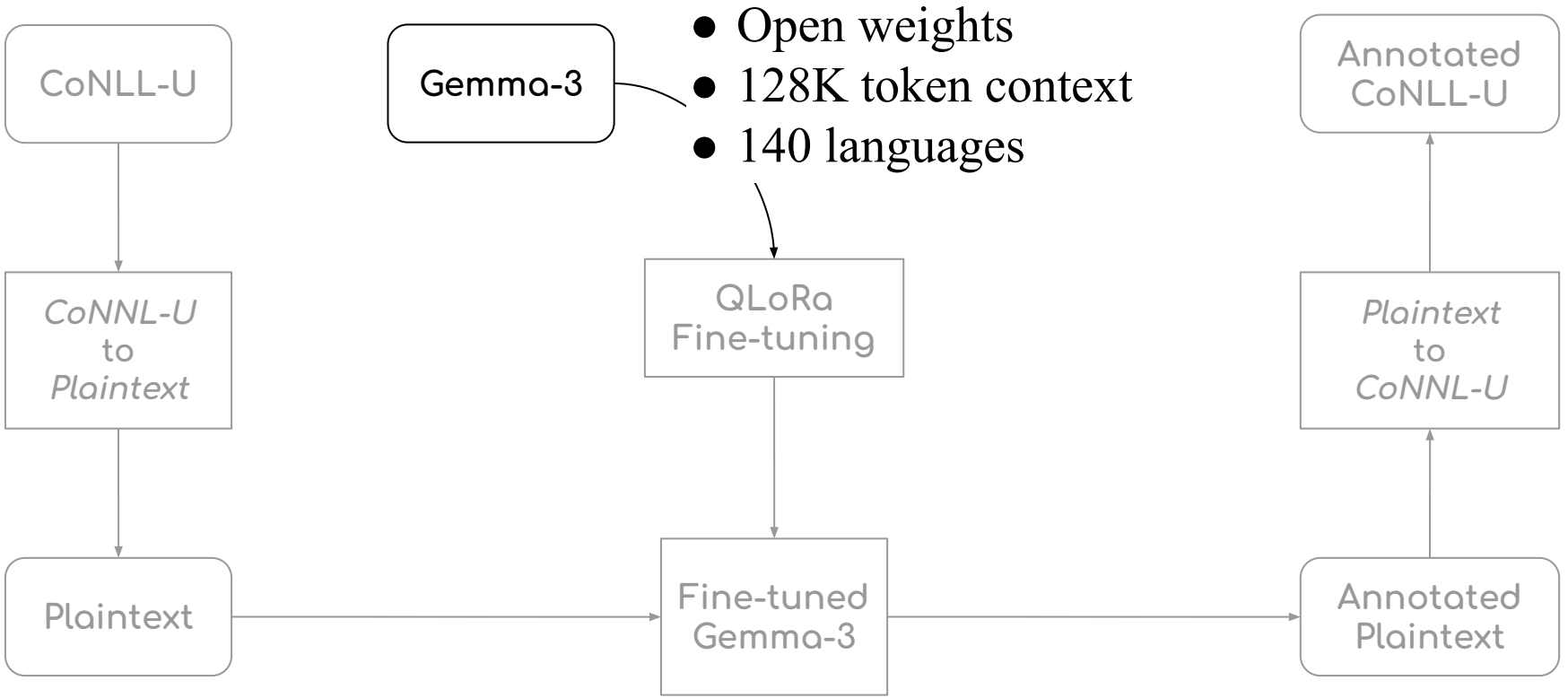
LLM Strategy: Fine-tuning



LLM Fine-tuning Details



Pre-trained LLM: Gemma-3 instruction-tuned (IT)



Gemma-3 Prompt Template

SYSTEM INSTRUCTION TEXT INPUT EXPECTED MODEL OUTPUT

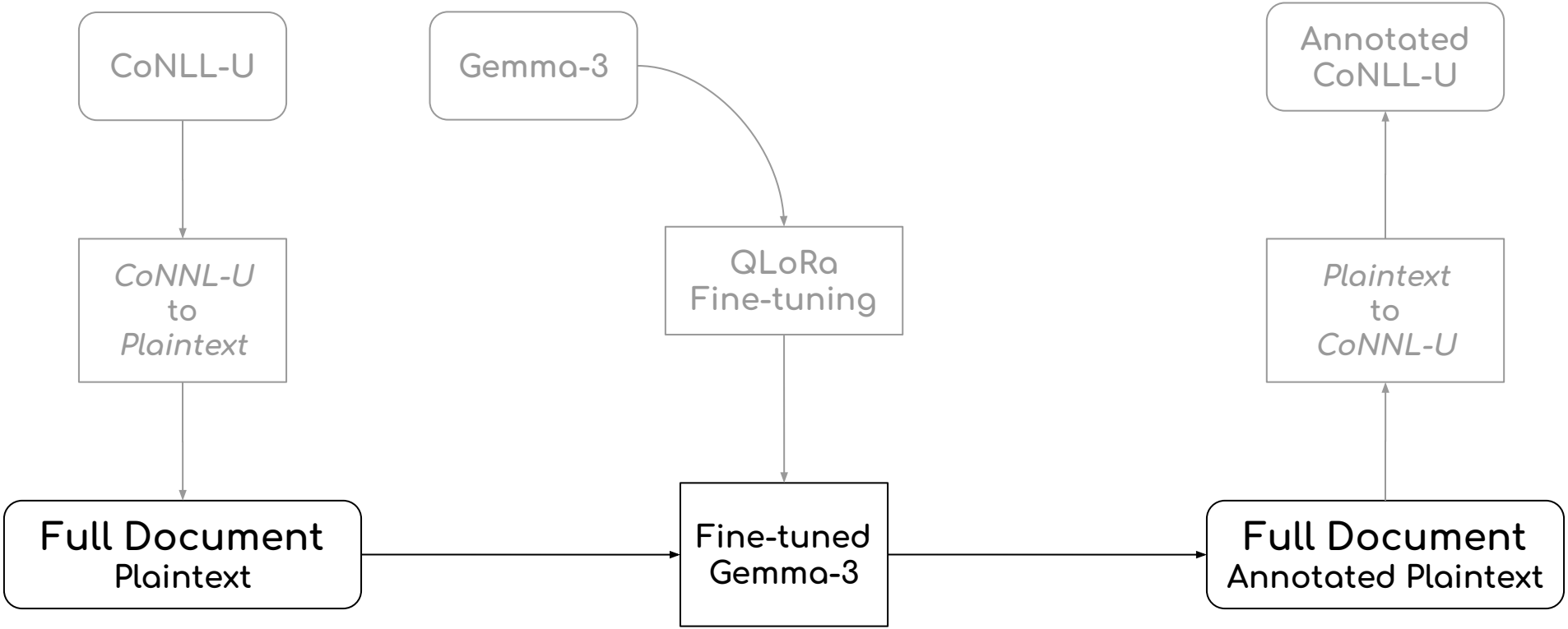
```
<start_of_turn>user
You are a linguist, expert in anaphora and coreference resolution.
Annotate in the input sentences which nouns, pronouns and other expressions
refer to the same entity.
Do only insert annotations. Do not insert extra linguistic material, nor
punctuation markers and do not delete elements from the input texts.

Input: *PLAINTEXT*
<end_of_turn>

<start_of_turn>model
*COREFERENCE ANNOTATED PLAINTEXT*
<end_of_turn>
```

Input format ?

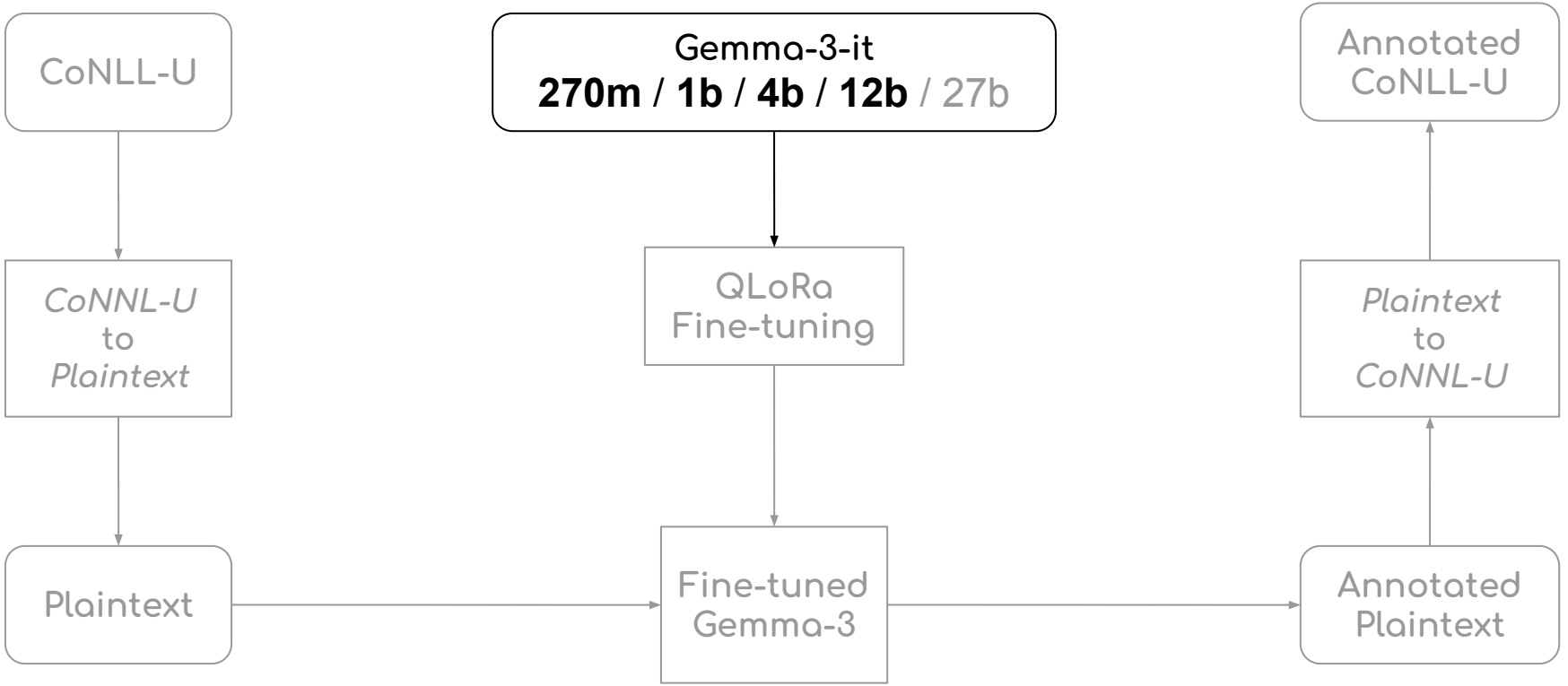
SYSTEM INSTRUCTION TEXT INPUT EXPECTED MODEL OUTPUT

```
<start_of_turn>user
You are a linguist, expert in anaphora and coreference resolution.
Annotate in the input sentences which nouns, pronouns and other expressions
refer to the same entity.
Do only insert annotations. Do not insert extra linguistic material, nor
punctuation markers and do not delete elements from the input texts.

Input: *PLAINTEXT*
<end_of_turn>

<start_of_turn>model
*COREFERENCE ANNOTATED PLAINTEXT*
<end_of_turn>
```

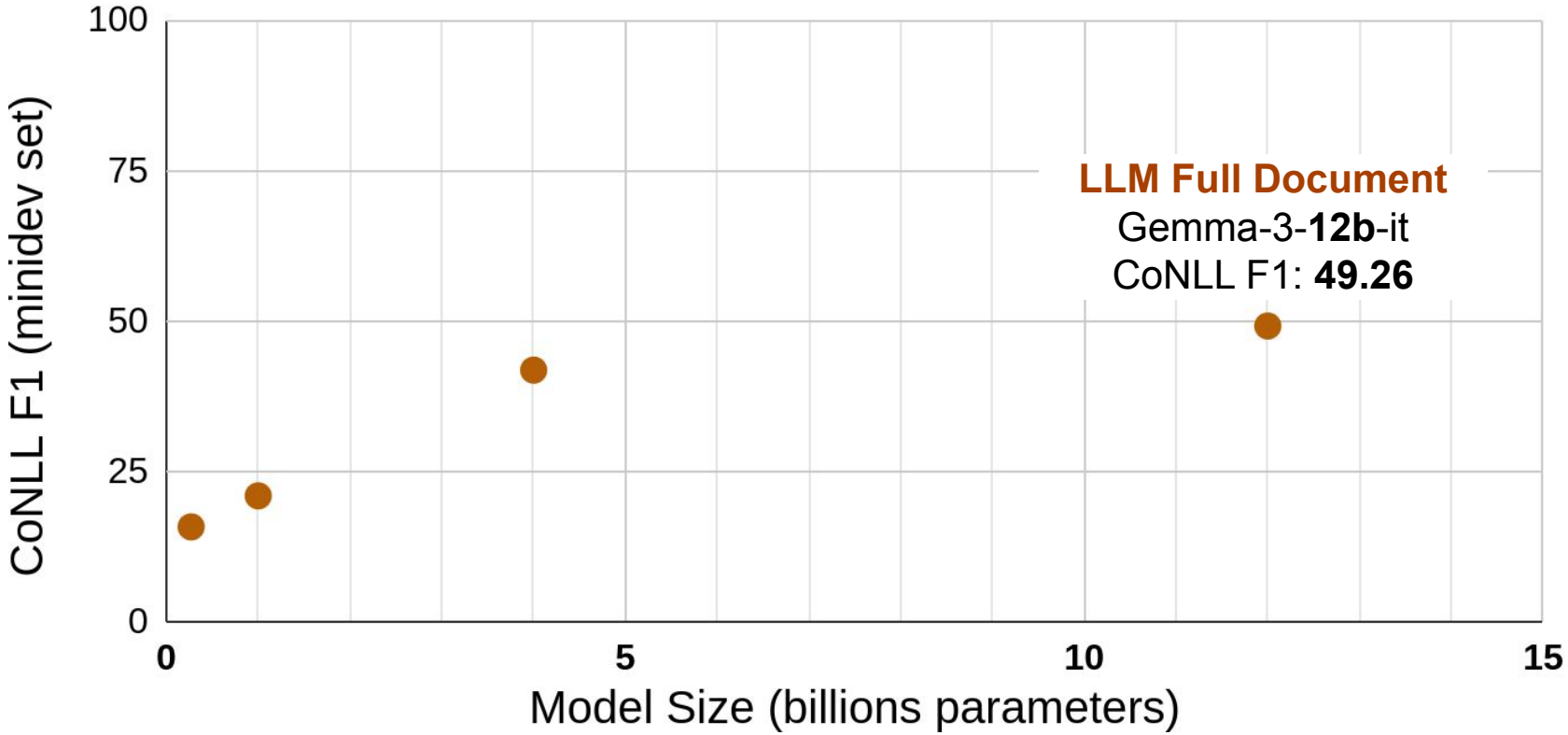
Full Document Annotation



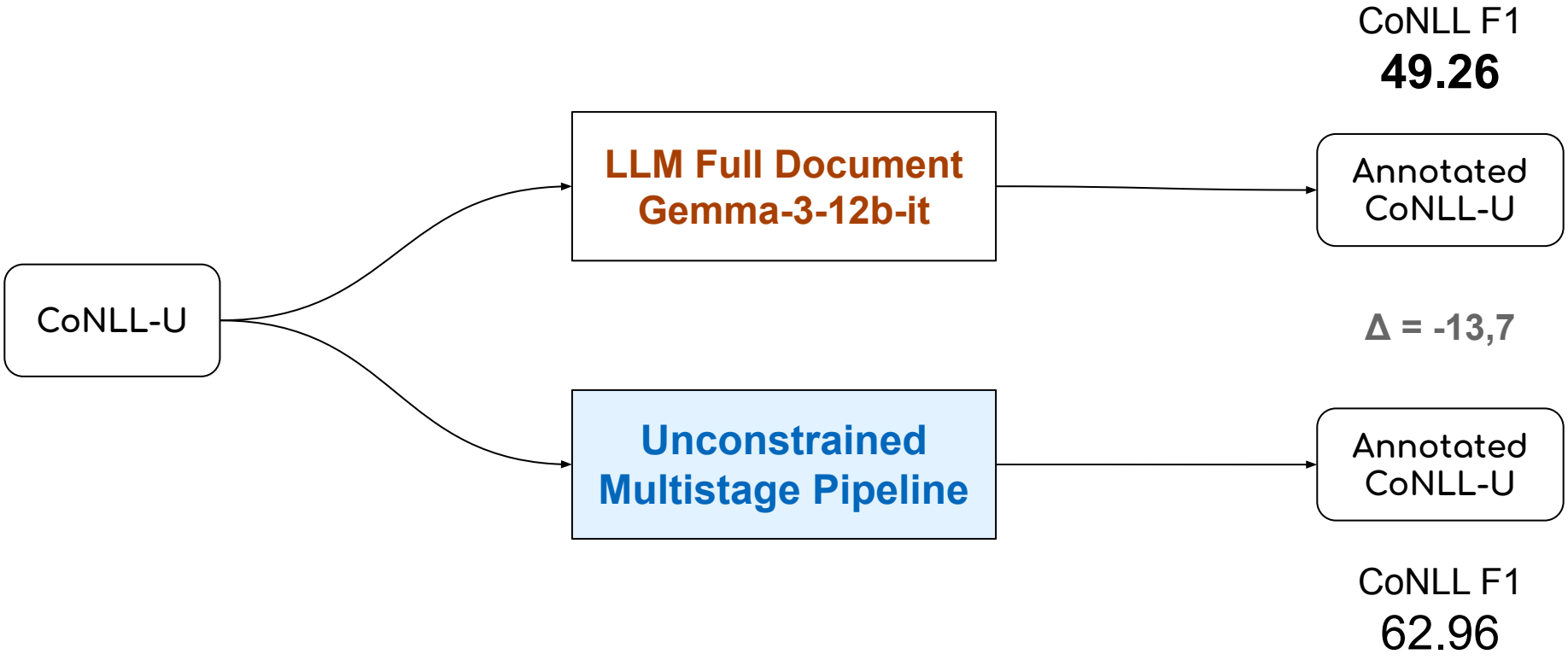
Full Document Annotation: Model Size Impact



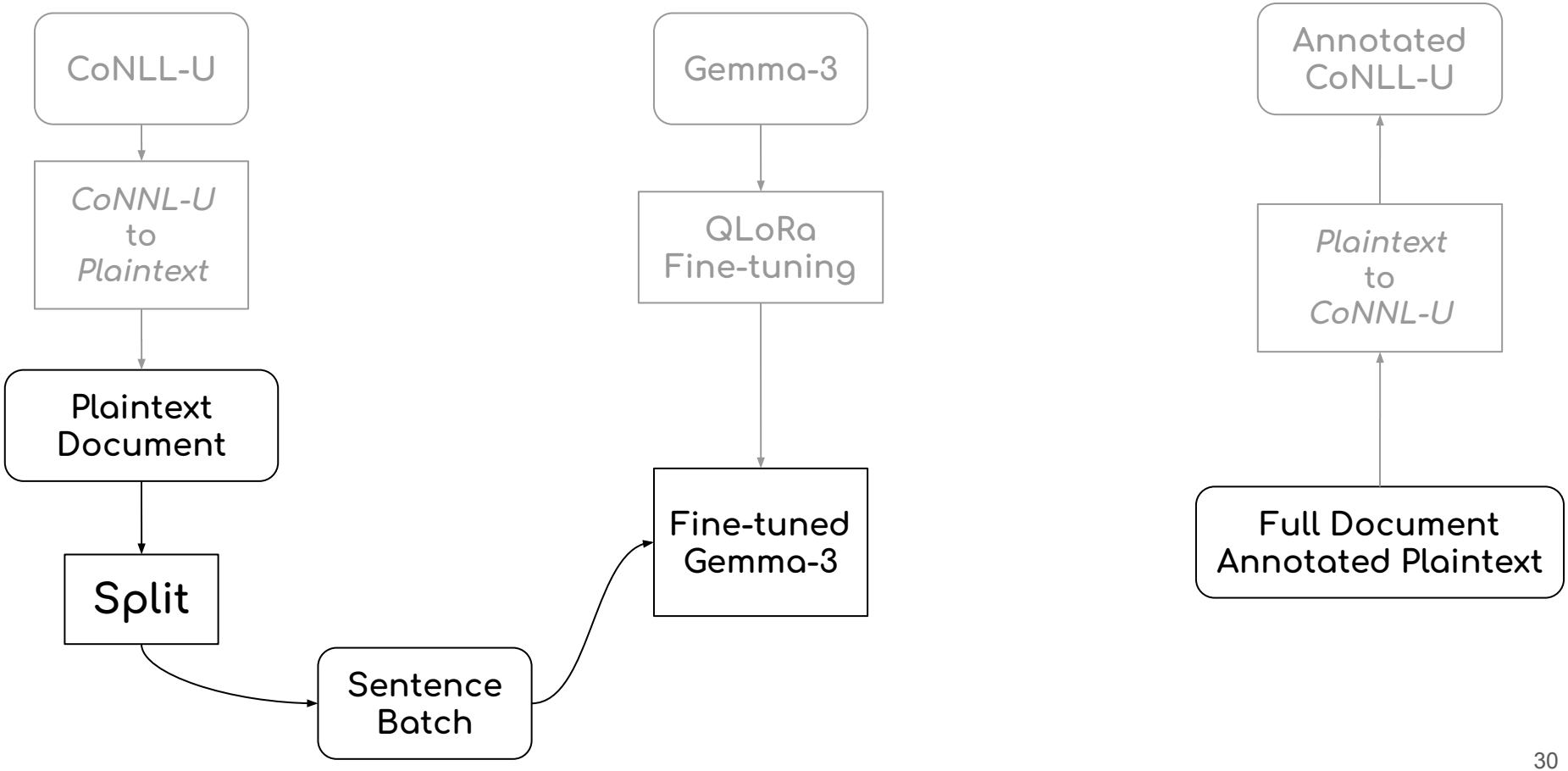
Full Document Annotation: Model Size Impact



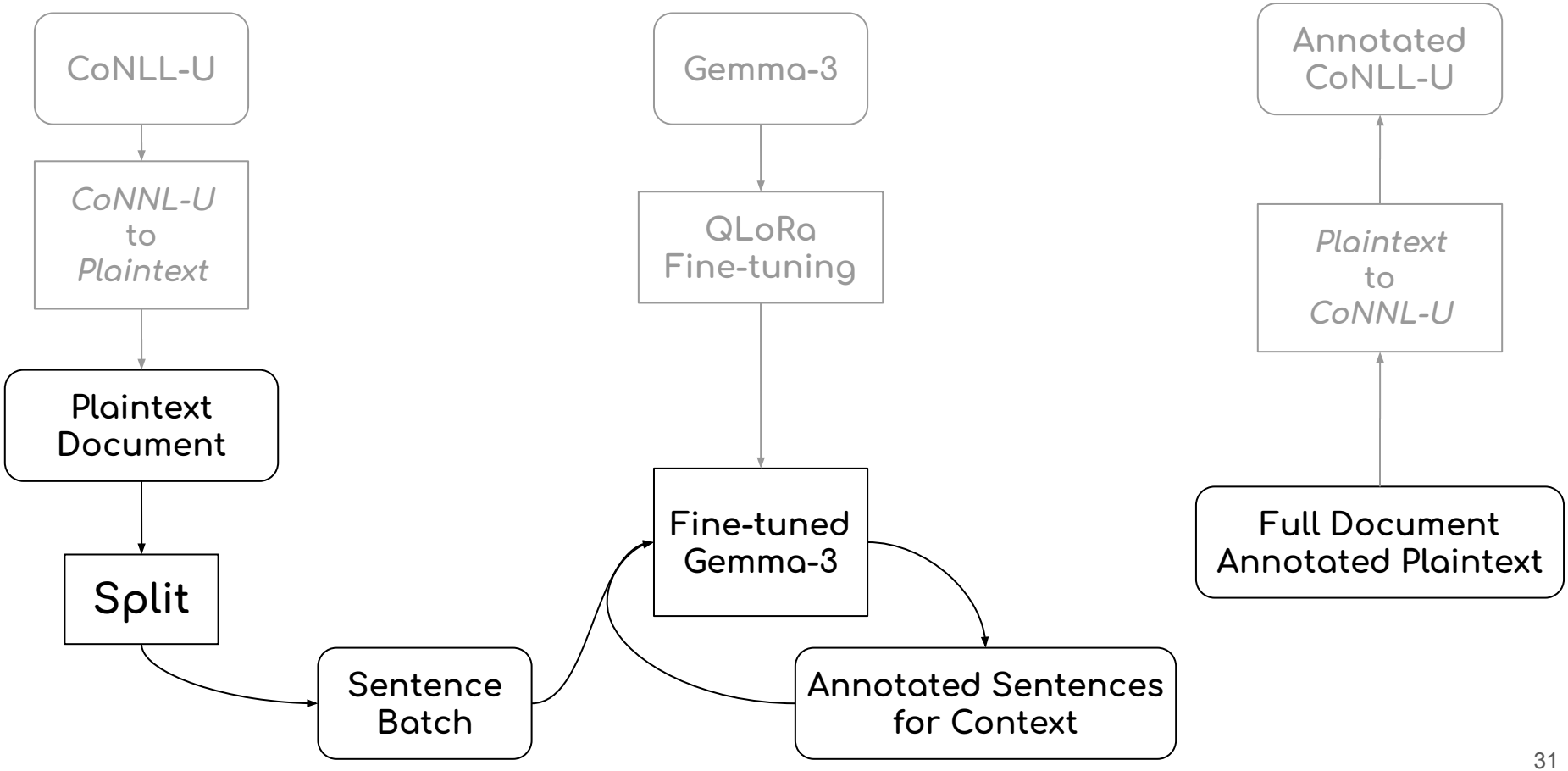
Unconstrained > LLM Full Document



Incremental Sentence Batch Annotation



Incremental Sentence Batch Annotation



Incremental Sentence Batch Annotation: Prompt Template

SYSTEM INSTRUCTION PREVIOUS CONTEXT TEXT INPUT EXPECTED MODEL OUTPUT

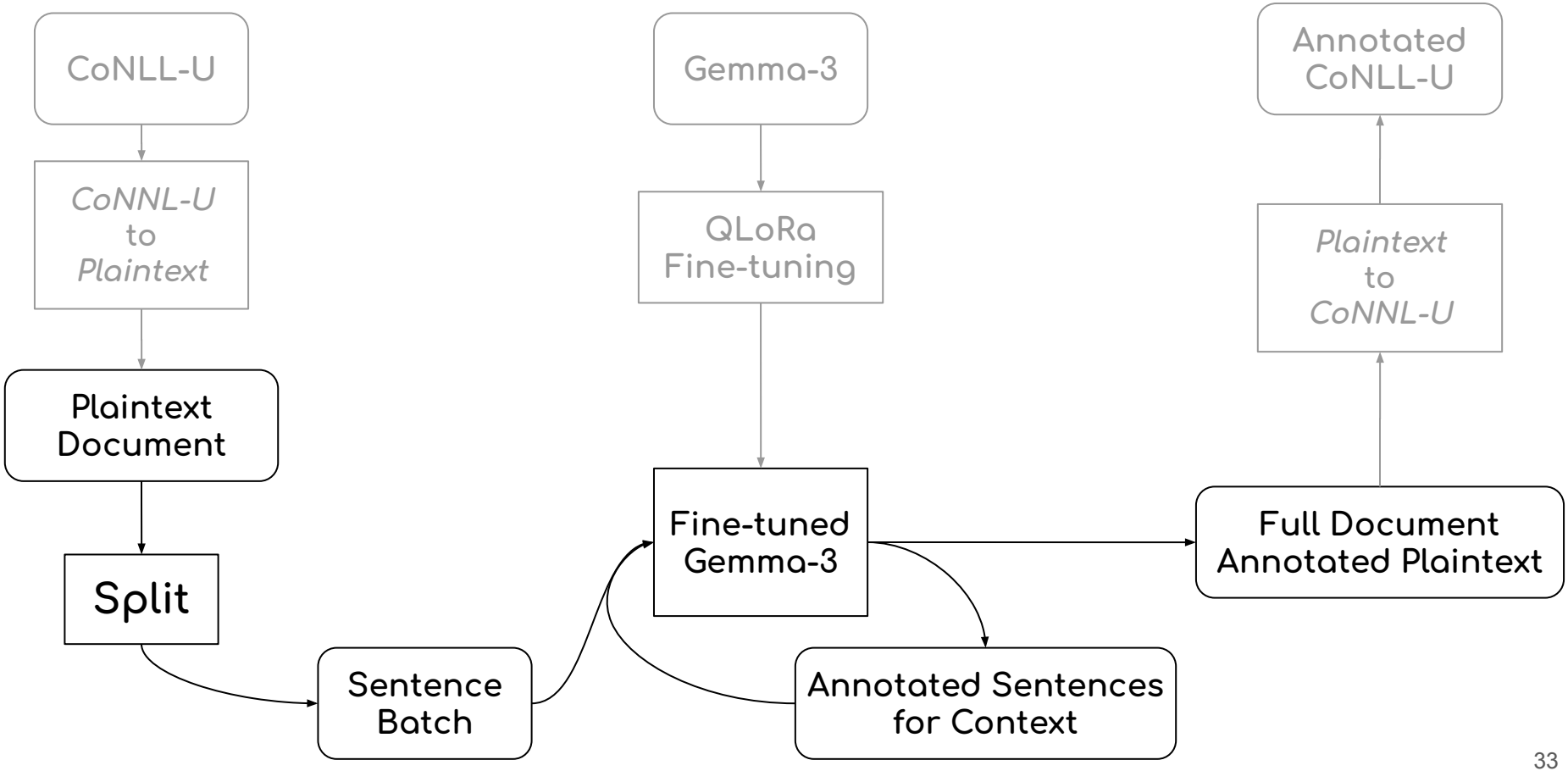
<start_of_turn>user
You are a linguist, expert in anaphora and coreference resolution.
Based on the previous context, annotate in the input sentences which nouns, pronouns and other expressions refer to the same entity.
Do only insert annotations. Do not insert extra linguistic material, nor punctuation markers and do not delete elements from the input texts.

Previous context: *ANNOTATED SENTENCES FROM PREVIOUS BATCHES*

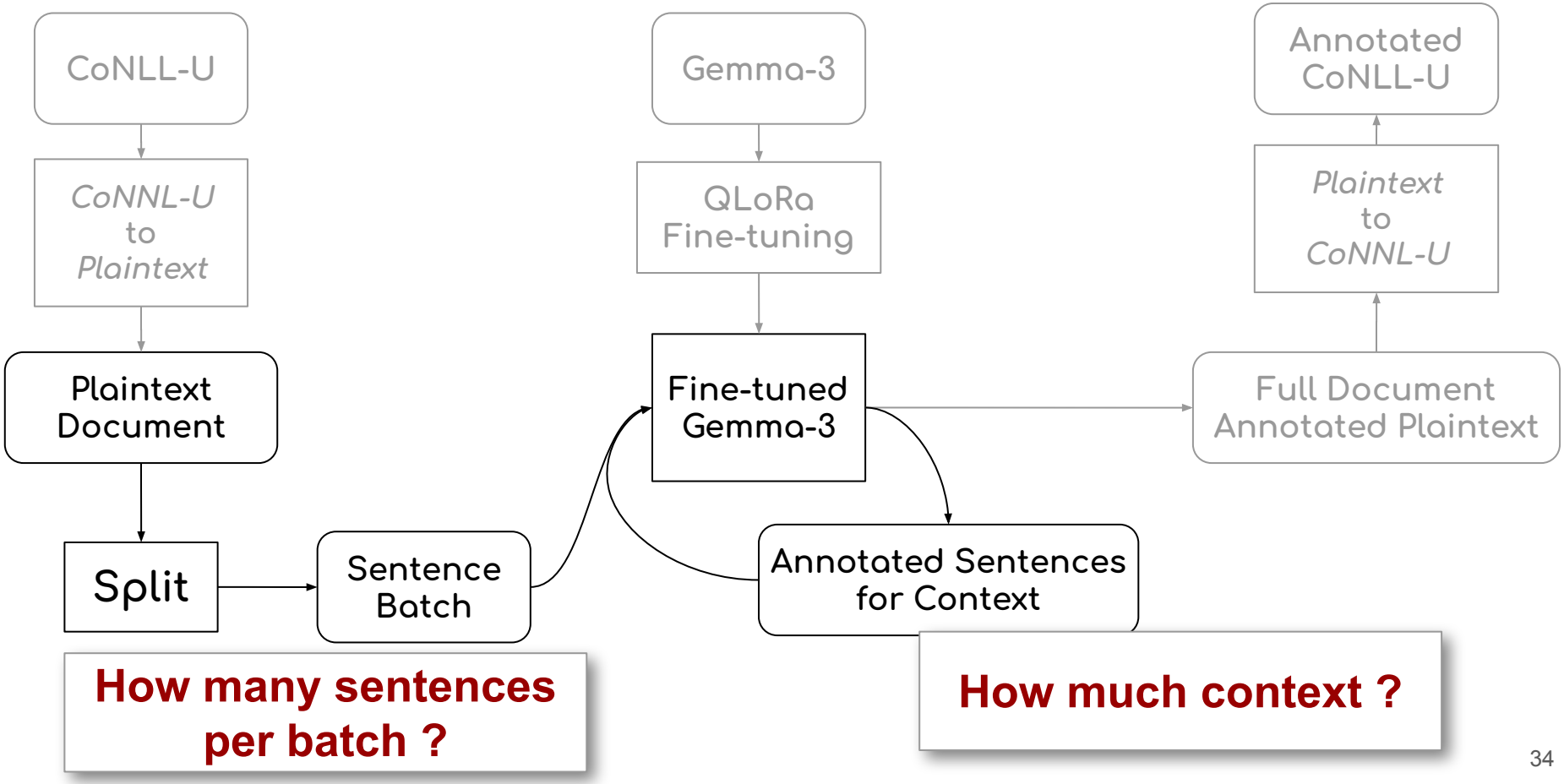
Input: *PLAINTEXT SENTENCE BATCH*
<end_of_turn>

<start_of_turn>model
COREFERENCE ANNOTATED SENTENCE BATCH
<end_of_turn>

Incremental Sentence Batch Annotation



Incremental Sentence Batch Annotation



Annotation Strategies

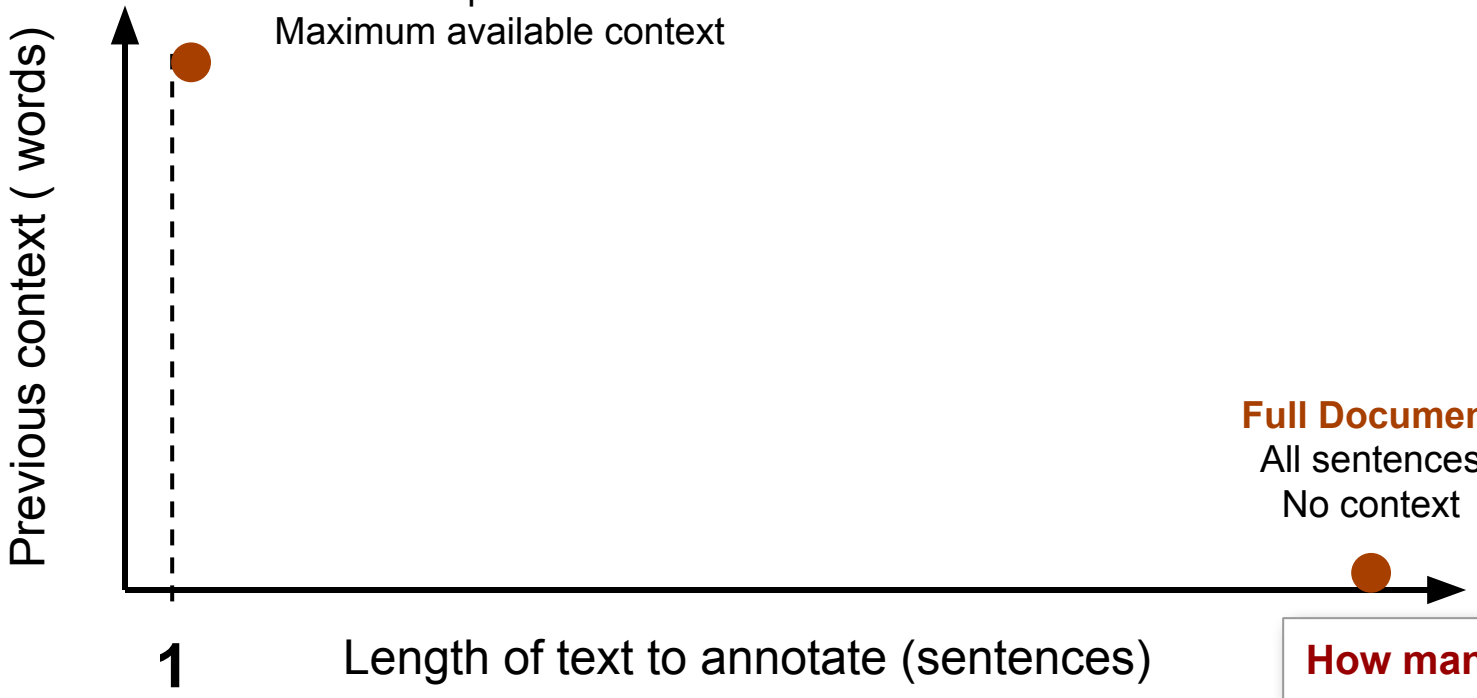
How much context ?

How many sentences per batch ?

Annotation Strategies

How much context ?

Incremental Sentence-by-Sentence
1 sentence per batch
Maximum available context



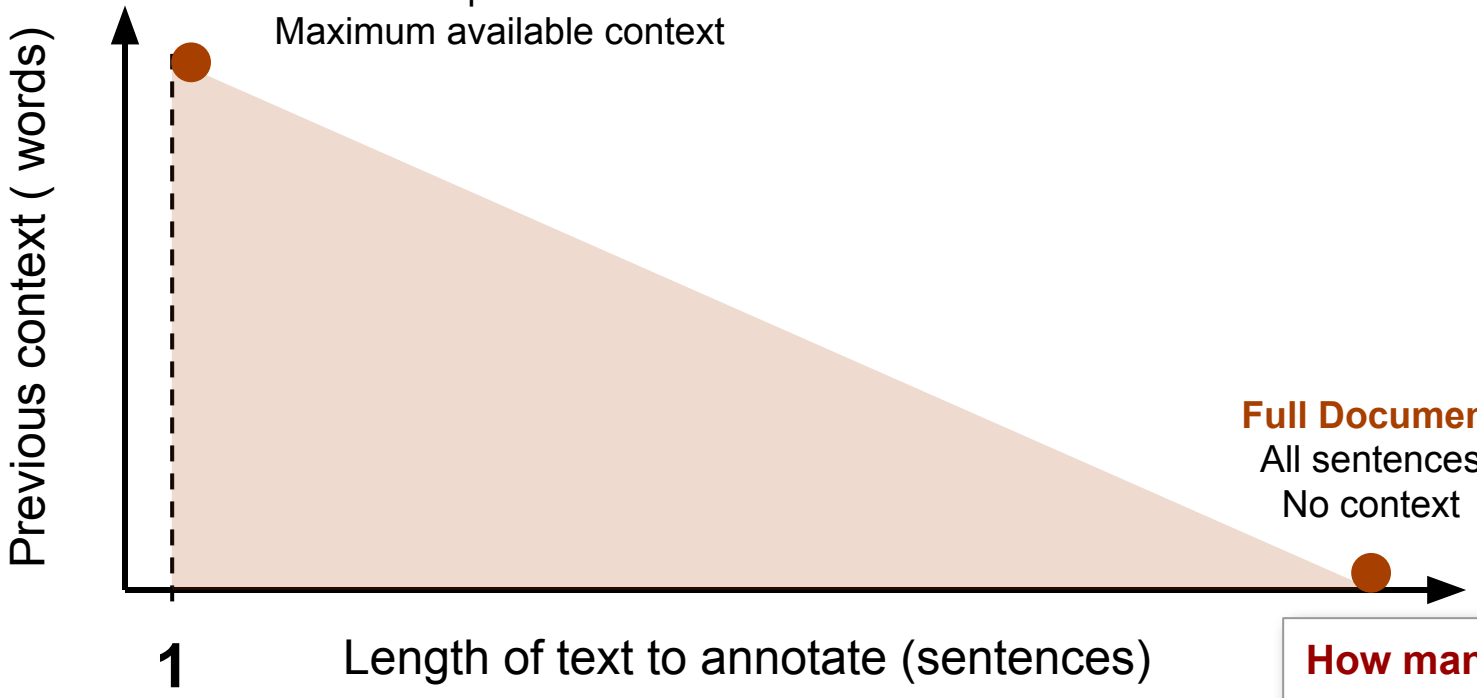
Full Document
All sentences
No context

How many sentences per batch ?

Space of Annotation Strategies

How much context ?

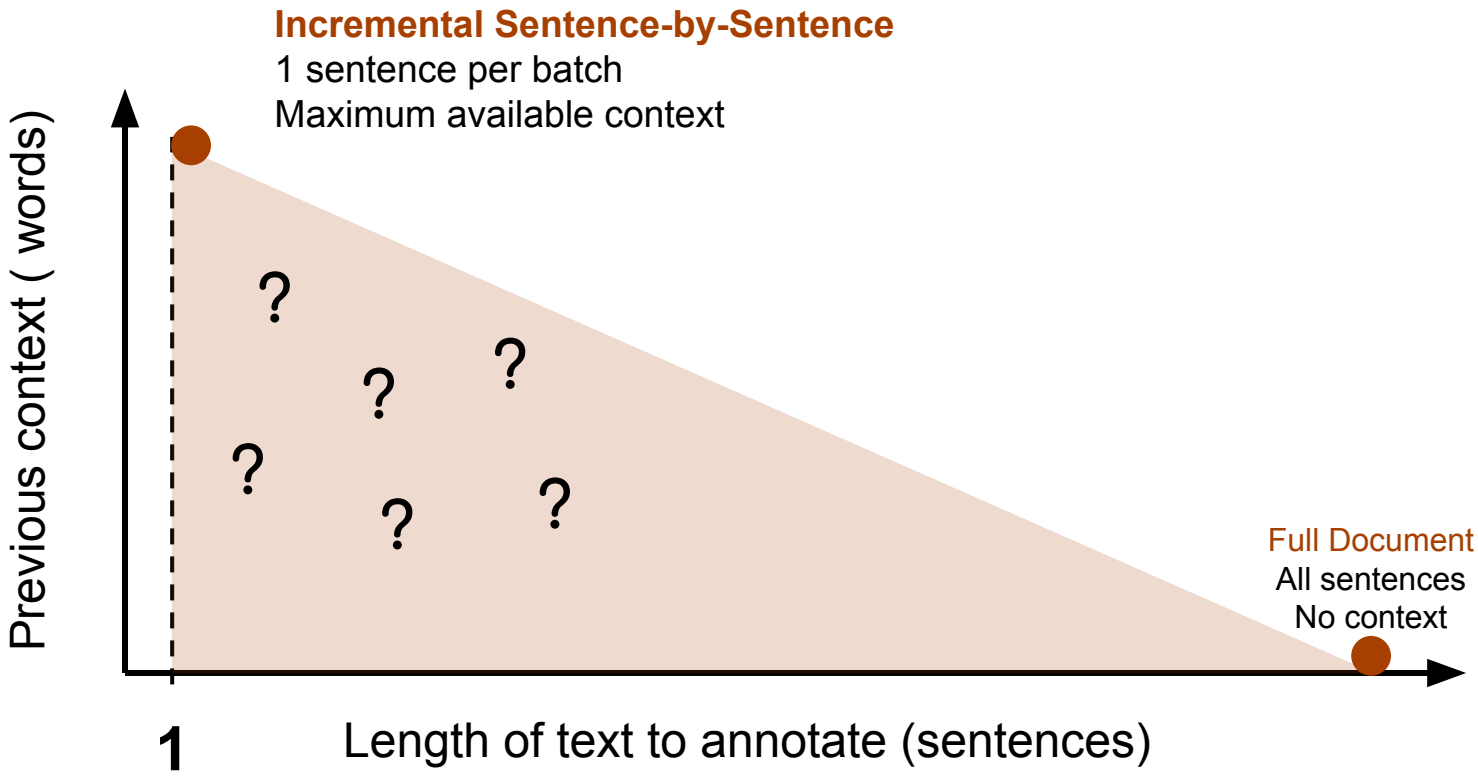
Incremental Sentence-by-Sentence
1 sentence per batch
Maximum available context



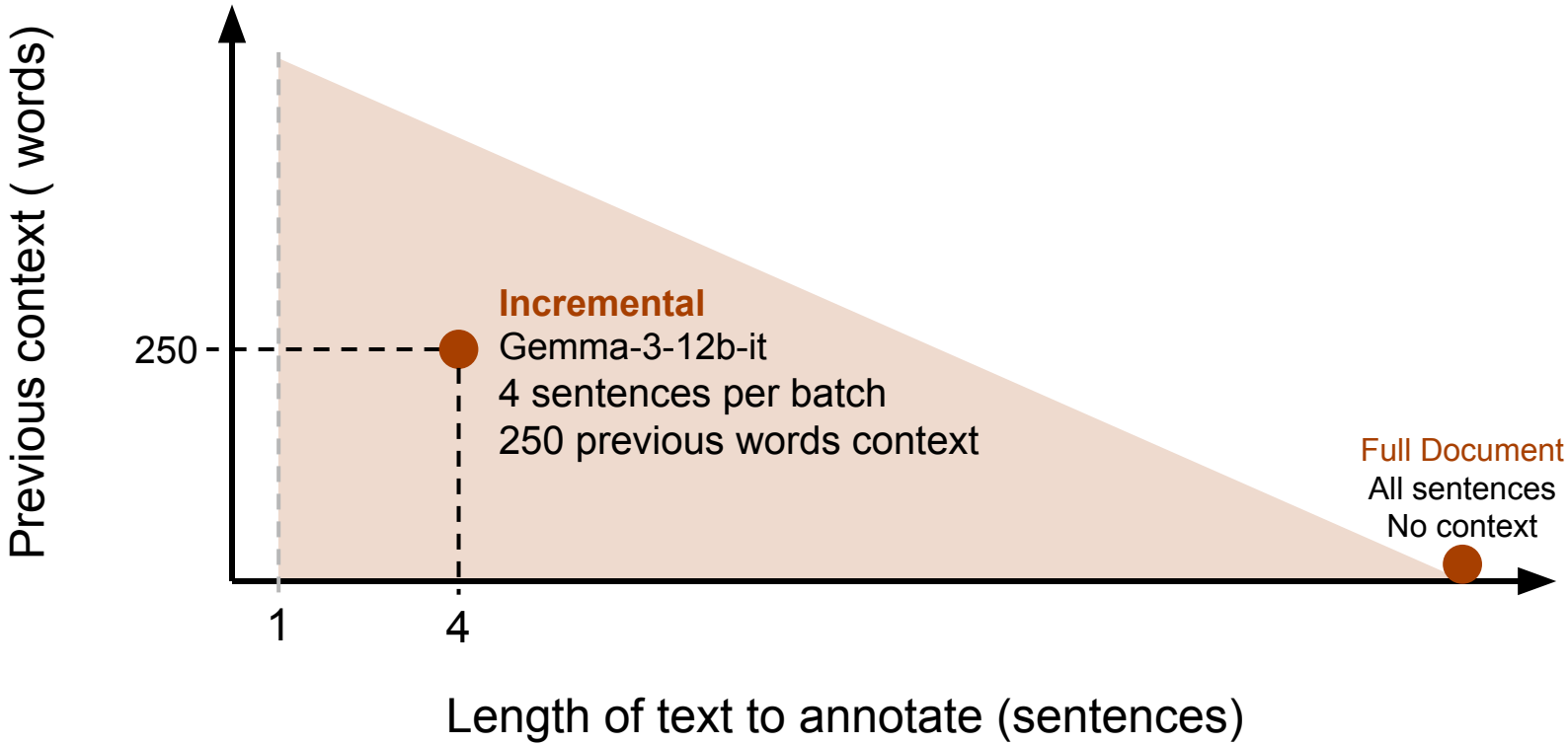
Full Document
All sentences
No context

How many sentences per batch ?

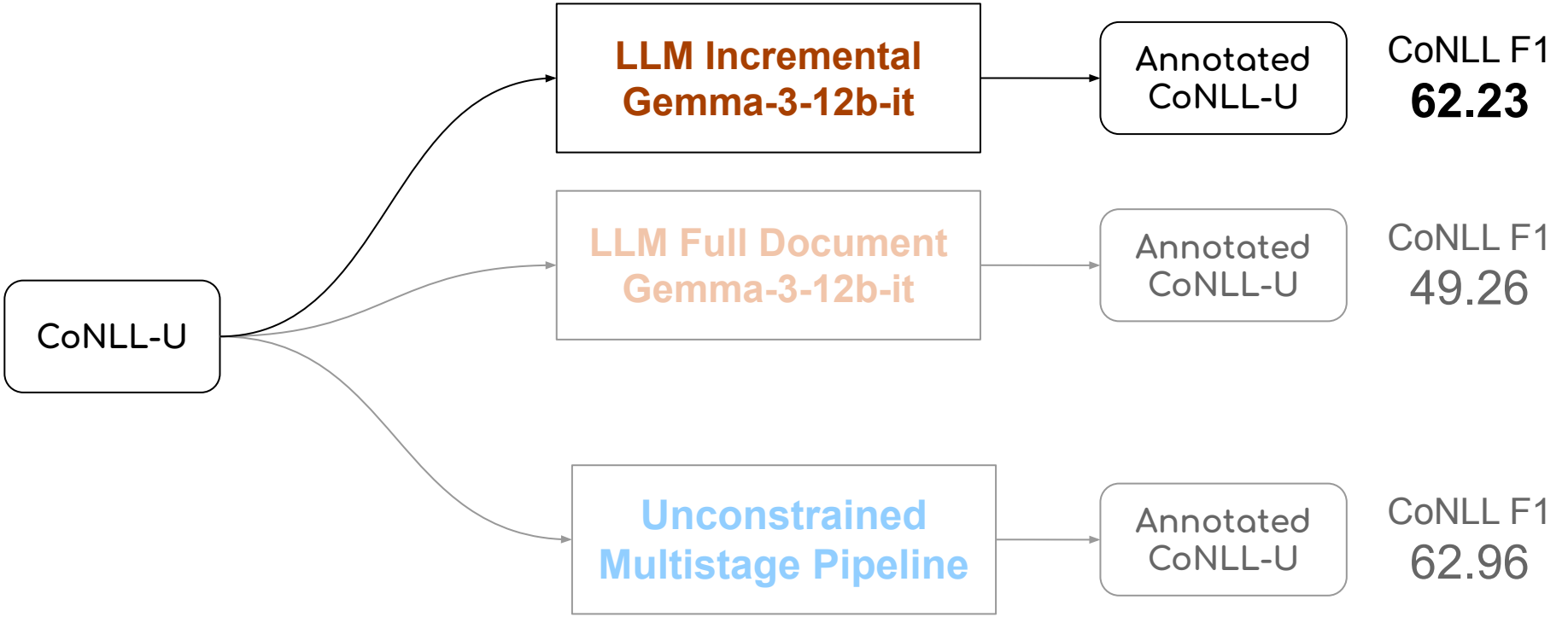
Space of Annotation Strategies: Exploratory Experiments



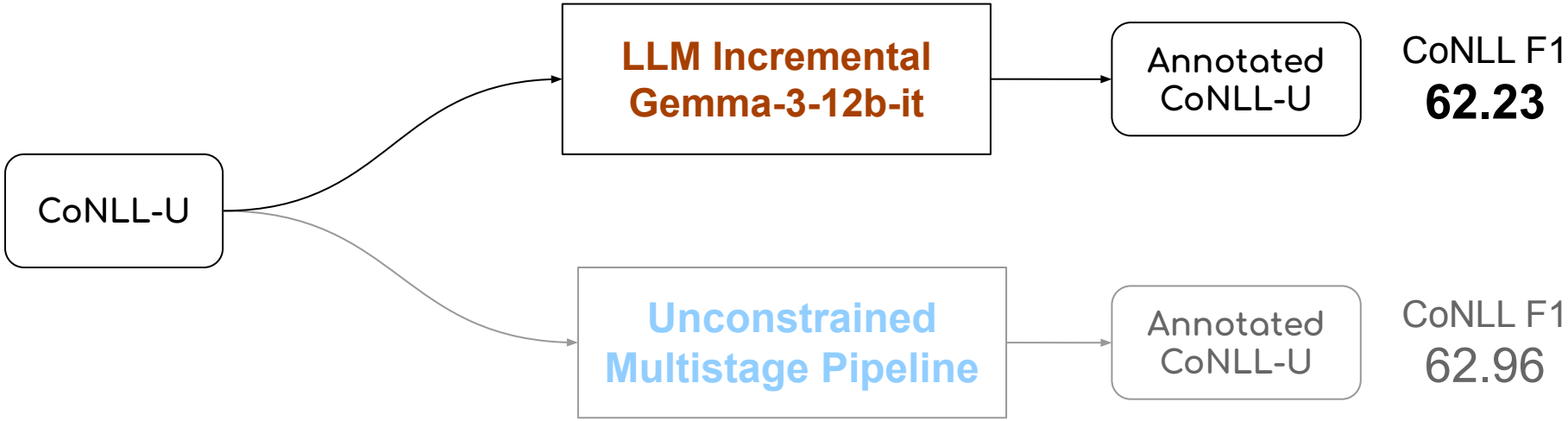
Space of Annotation Strategies: Selected Parameters



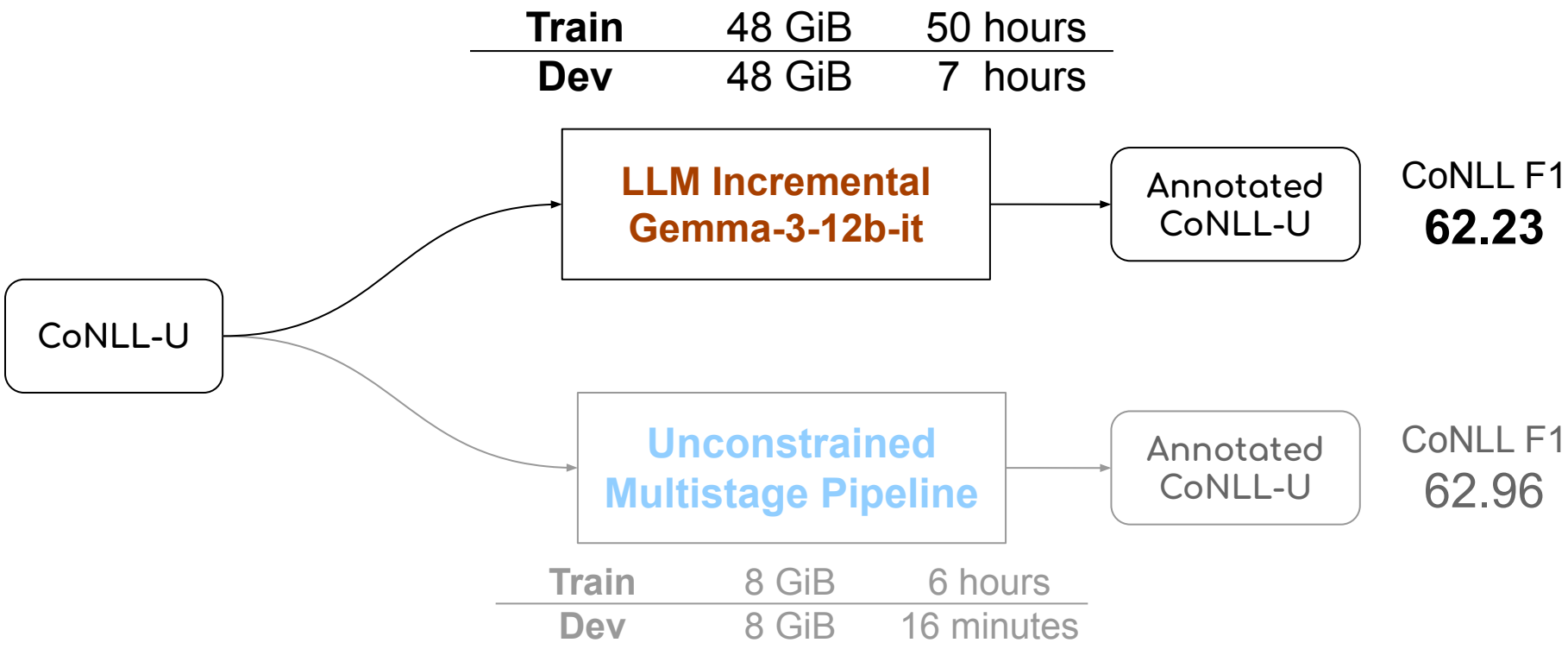
LLM Incremental: Results



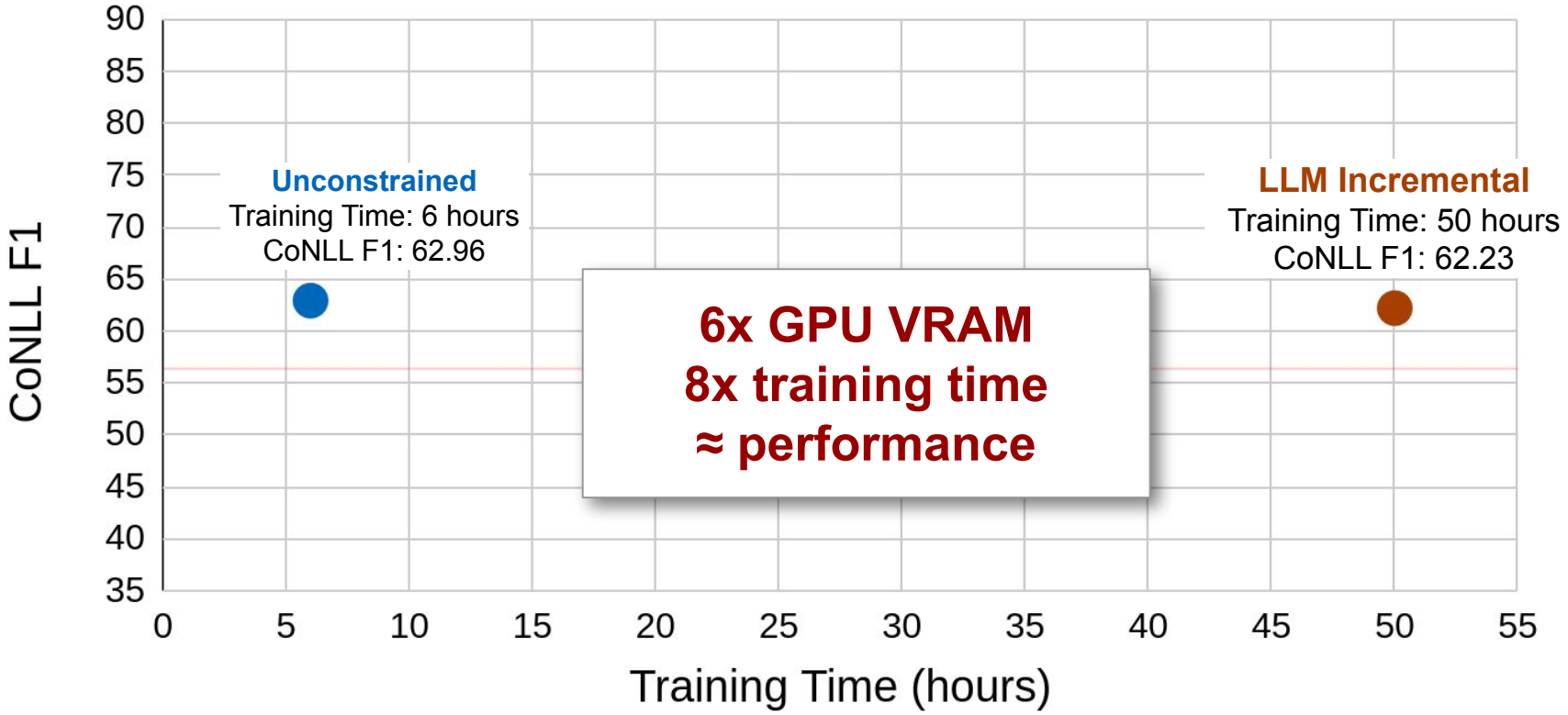
LLM Incremental: Best Model



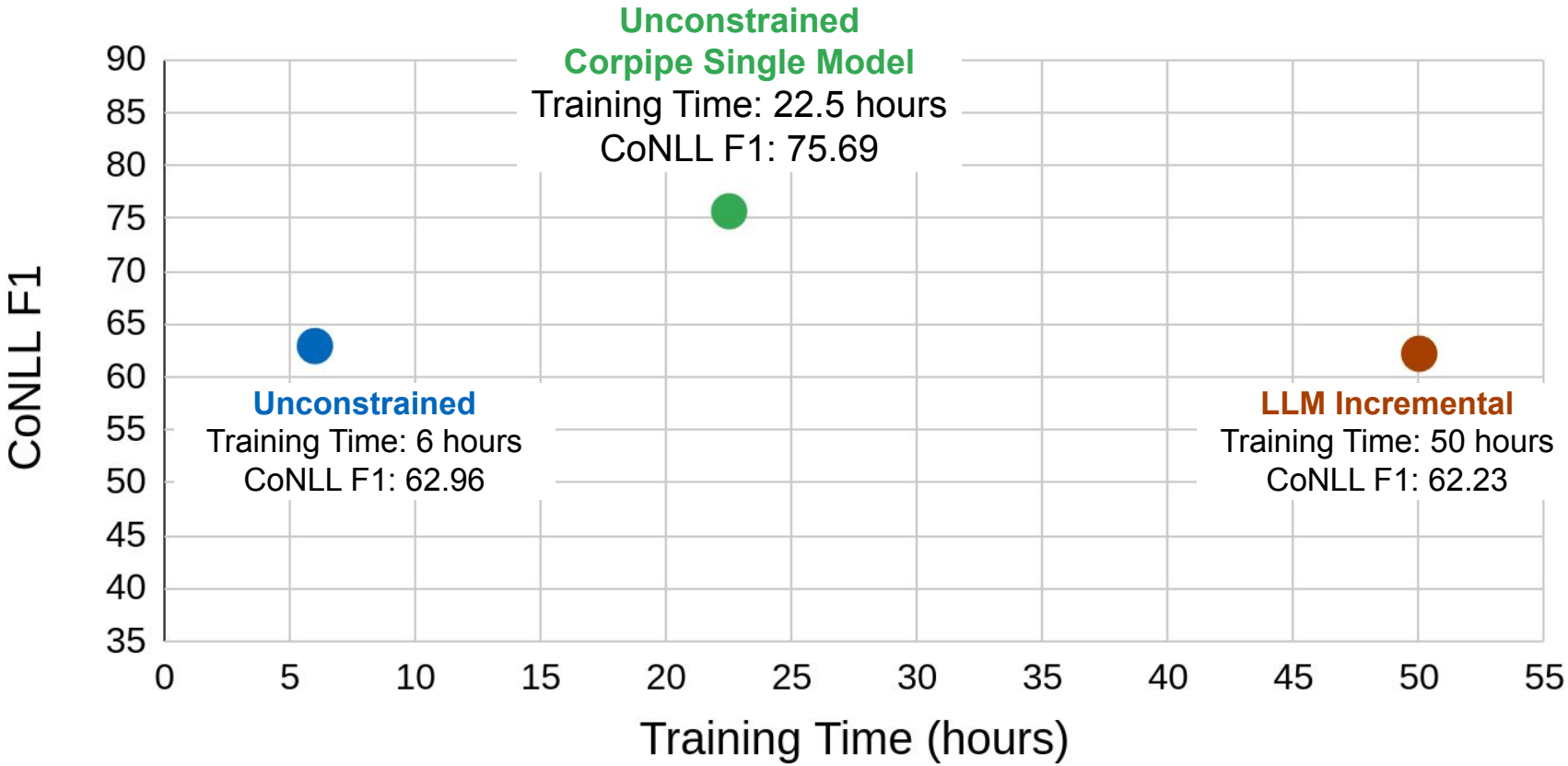
LLM Incremental: Ressources



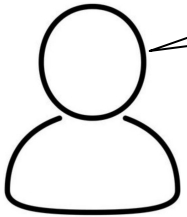
Comparison of Models



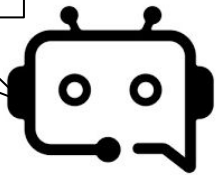
Comparison: Best Unconstrained Single Model Solution



Should We Transform Coreference Resolution into a Text Generation Task?



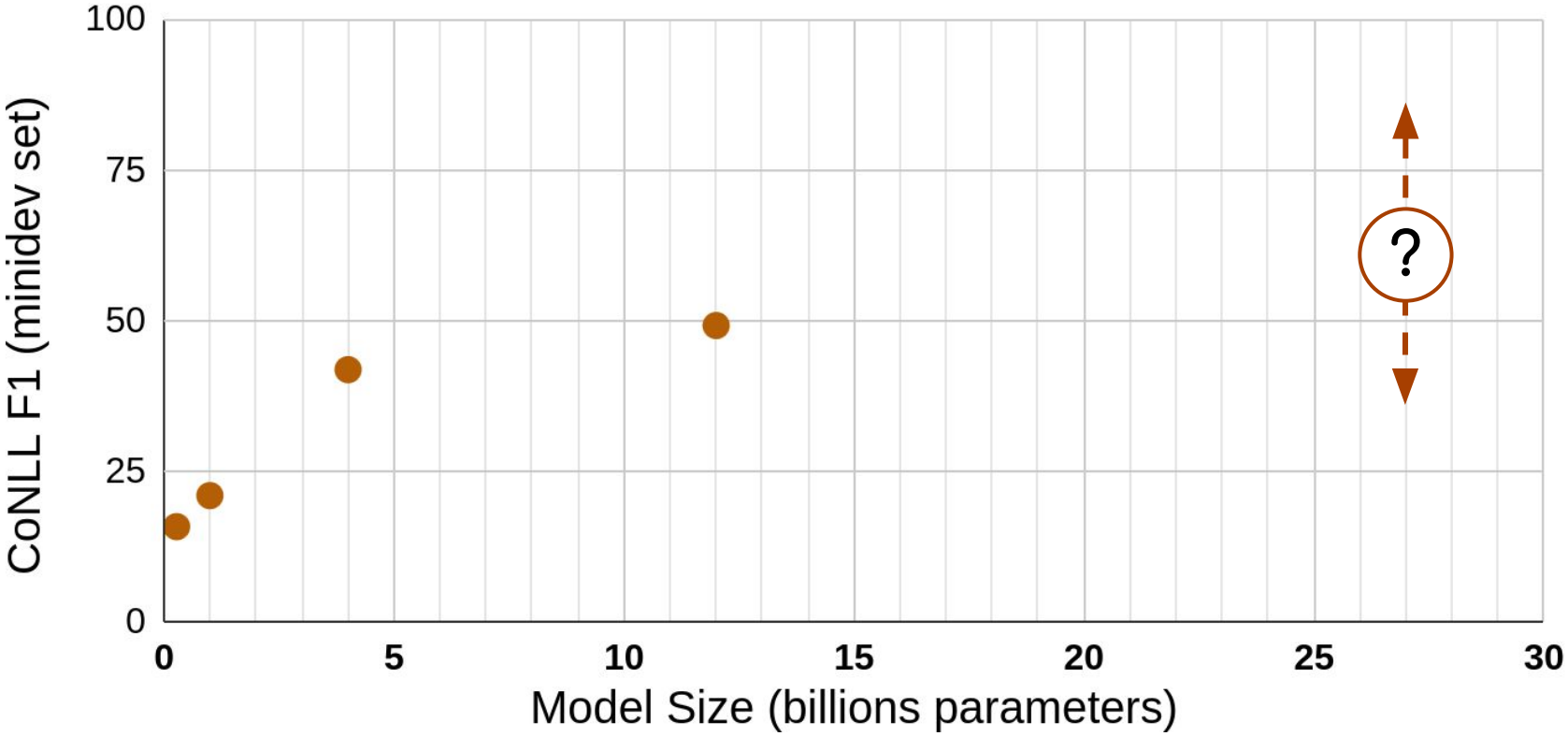
Not yet.



LLM Coreference Resolution: Perspectives

1. Model Size Increase
2. Coreference ID Tracking
3. Plaintext Format

Perspectives: Model Size Increase



LLM Coreference Resolution: Perspectives

1. Model Size Increase
2. Annotated Entities Tracking
3. Plaintext Format

Perspectives: Annotated Entities Tracking

SYSTEM INSTRUCTION PREVIOUS CONTEXT TEXT INPUT EXPECTED MODEL OUTPUT

```
<start_of_turn>user
You are a linguist, expert in anaphora and coreference resolution.
Based on the previous context, annotate in the input sentences which nouns,
pronouns and other expressions refer to the same entity.
Do only insert annotations. Do not insert extra linguistic material, nor
punctuation markers and do not delete elements from the input texts.

Previous context: *ANNOTATED SENTENCES FROM PREVIOUS BATCHES*

Entity Tracker: "Alice's sister"[e1], "Alice"[e2], "The White Rabbit"[e3]

Input: *PLAINTEXT SENTENCE BATCH*
<end_of_turn>

<start_of_turn>model
*COREFERENCE ANNOTATED SENTENCE BATCH*
<end_of_turn>
```

LLM Coreference Resolution: Perspectives

1. Model Size Increase
2. Coreference ID Tracking
3. Plaintext Format Modification

Perspectives: Plaintext Format Modification

Down the**|[e1** Rabbit-Hole**|e1]** Alice**|[e2]** was beginning to get very tired of sitting by her**|[e2],[e3** sister**|e3]** on the **|[e4** bank**|e4]** , and of having nothing to do : once or twice she**|[e2]** had peeped into the book her**|[e2],[e3** sister**|e3]** was reading.

Down **<e1>**the Rabbit-Hole**</e1>** **<e2>**Alice**</e2>** was beginning to get very tired of sitting by **<e3><e2>**her**</e2>** sister**</e3>** on **<e4>**the bank**</e4>** , and of having nothing to do : once or twice **<e2>**she**</e2>** had peeped into thebook **<e3><e2>**her**</e2>** sister**</e3>** was reading.

Alternative tagging scheme inspired by markup languages like HTML or XML that tokenizers and LLMs might be more familiar with.

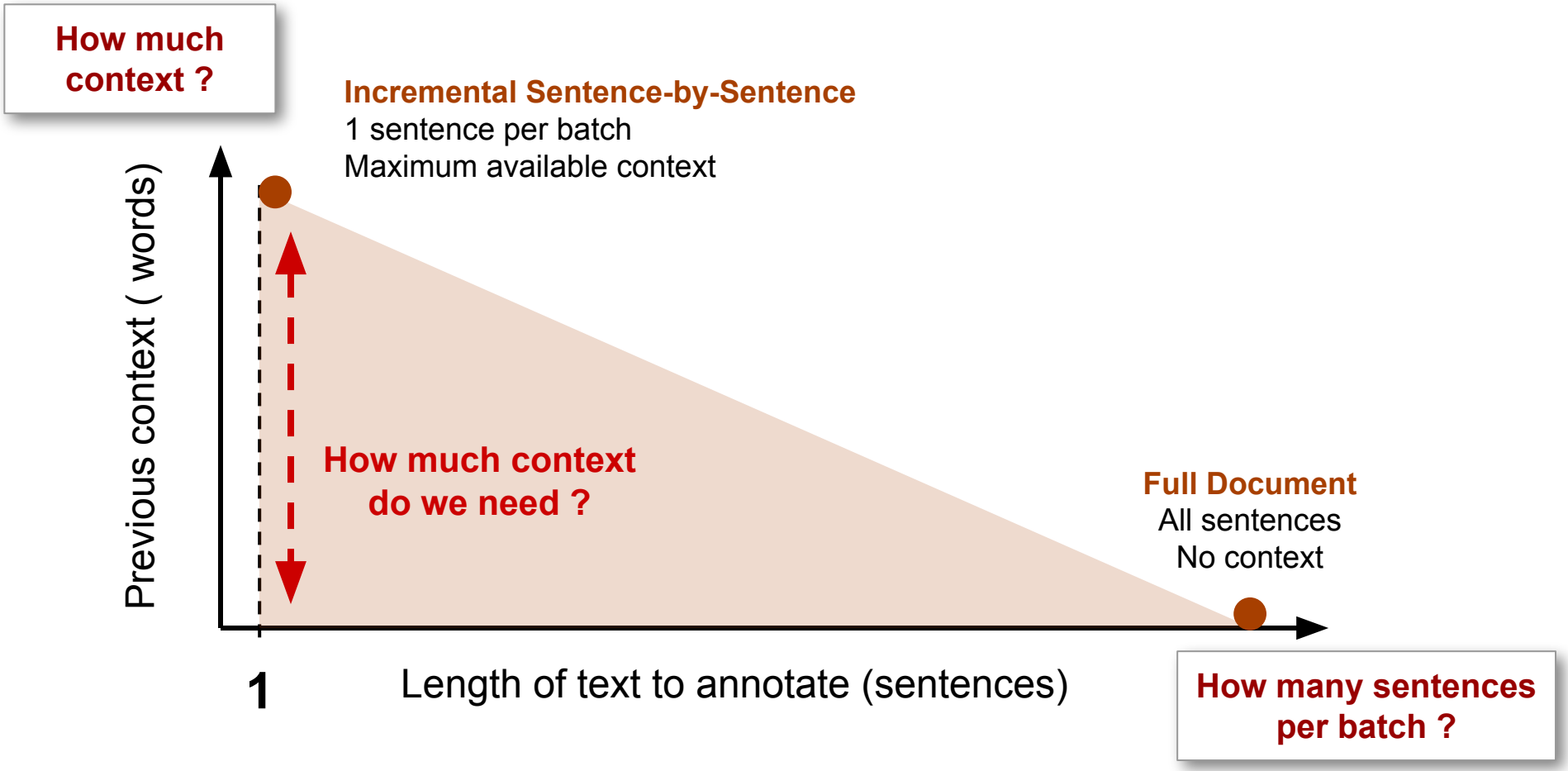
Perspectives: Coreference ID Tracking

1. Model Size Increase
2. Coreference ID Tracking
3. Plaintext Format
4. **Other Suggestions ?**

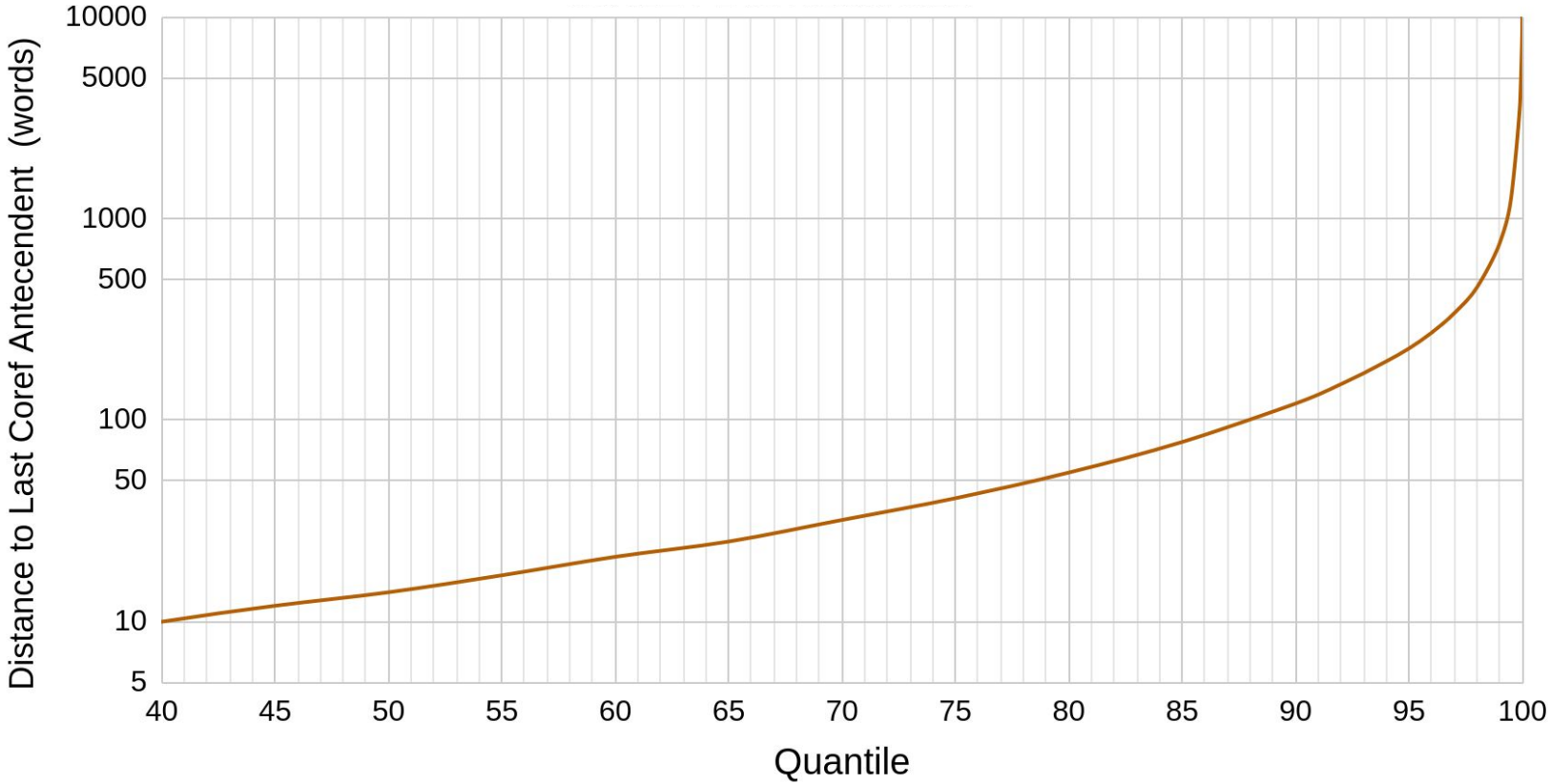
Thank you

Antoine BOURGOIS
Lattice (CNRS – École Normale Supérieure – Université Sorbonne Nouvelle), Paris, France
antoine.bourgois@ens.psl.eu

Additional Material: Space of Annotation Strategies



Additional Material: Distance to Last Coreferential Antecedent

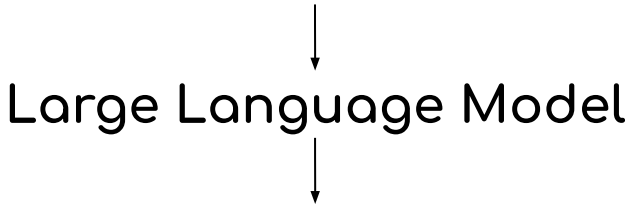


Additional Material: Incremental approach

PREVIOUS CONTEXT RAW TEXT INPUT EXPECTED MODEL OUTPUT

STEP 1

[None] CHAPTER I. Down the Rabbit-Hole Alice was beginning to get very tired of sitting by her sister on the bank , and of having nothing to do : once or twice she had peeped into the book her sister was reading , but it had no pictures or conversations in it , ‘ and what is the use of a book , ’ thought Alice ‘ without pictures or conversations ? ’



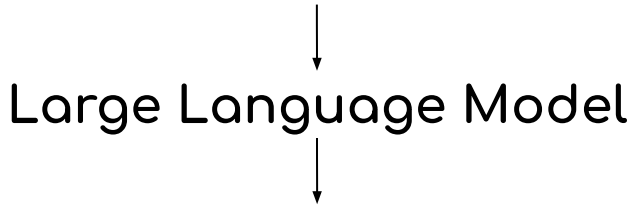
CHAPTER I. Down the|[e1 Rabbit-Hole|e1] Alice|[e2] was beginning to get very tired of sitting by her|[e2],[e3 sister|e3] on the|[e4 bank|e4] , and of having nothing to do : once or twice she|[e2] had peeped into the book her|[e2],[e3 sister|e3] was reading , but it had no pictures or conversations in it , ‘ and what is the use of a book , ’ thought Alice|[e2] ‘ without pictures or conversations ? ’

Additional Material: Incremental approach

PREVIOUS CONTEXT RAW TEXT INPUT EXPECTED MODEL OUTPUT

STEP 2

CHAPTER I. Down the|[e1 Rabbit-Hole|e1] Alice|[e2] was beginning to get very tired of sitting by her|[e2],[e3 sister|e3] on the|[e4 bank|e4] , and of having nothing to do : once or twice she|[e2] had peeped into the book her|[e2],[e3 sister|e3] was reading , but it had no pictures or conversations in it , ‘ and what is the use of a book , ’ thought Alice|[e2] ‘ without pictures or conversations ? ’ So she was considering in her own mind (as well as she could , for the hot day made her feel very sleepy and stupid) , whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies , when suddenly a White Rabbit with pink eyes ran close by her .



So she|[e2] was considering in her|[e2] own mind (as well as she|[e2] could , for the hot day made her|[e2] feel very sleepy and stupid) , whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies , when suddenly a|[e5 White Rabbit with pink eyes|e5] ran close by her|[e2] .

Additional Material: All experiments

WordsContext	Sentence Batch	Model	Model Size (B)	CoNLL F1
0	max.	gemma-3-270m-it	0.27	15.91
0	max.	gemma-3-1b-it	1	21.06
100	4	gemma-3-270m-it	0.27	34.56
max.	8	gemma-3-1b-it	1	38.16
0	max.	gemma-3-4b-it	4	41.95
1000	8	gemma-3-1b-it	1	45.03
250	4	gemma-3-1b-it	1	47.53
0	max.	gemma-3-12b-it	12	49.26
100	8	gemma-3-1b-it	1	50.32
100	4	gemma-3-1b-it	1	50.89
100	4	gemma-3-4b-it	4	52.52
100	4	gemma-3-12b-it	12	56.32
250	8	gemma-3-4b-it	4	56.93
250	4	gemma-3-12b-it	12	59.34
250	10	gemma-3-12b-it	12	61.46
300	12	gemma-3-12b-it	12	61.55
250	8	gemma-3-12b-it	12	62.56
250	4	gemma-3-12b-it	12	62.65
250	10	gemma-3-27b-it	27	62.85