Langage, IA générative et robotique

ANITI Days 2024

25 novembre 2024

Learning by seeing and discussing: actions with collaborative conversation.

Akshay Chaturvedi & Kate Thompson

November 25, 2024

ANITI

How can we construct models that use conversational and physical context (which would underpin collaborative agents) ?

Experiment with generative approach to:

1. retrieve conversational structure
2. predict actions from instructions

ANITI

How can we construct models that use conversational and physical context (which would underpin collaborative agents) ?

Experiment with generative approach to:

1. retrieve conversational structure → discourse parser
2. predict actions from instructions → instruction-action model
3. how to leverage discourse information for action prediction

The Minecraft Dialogue Corpus

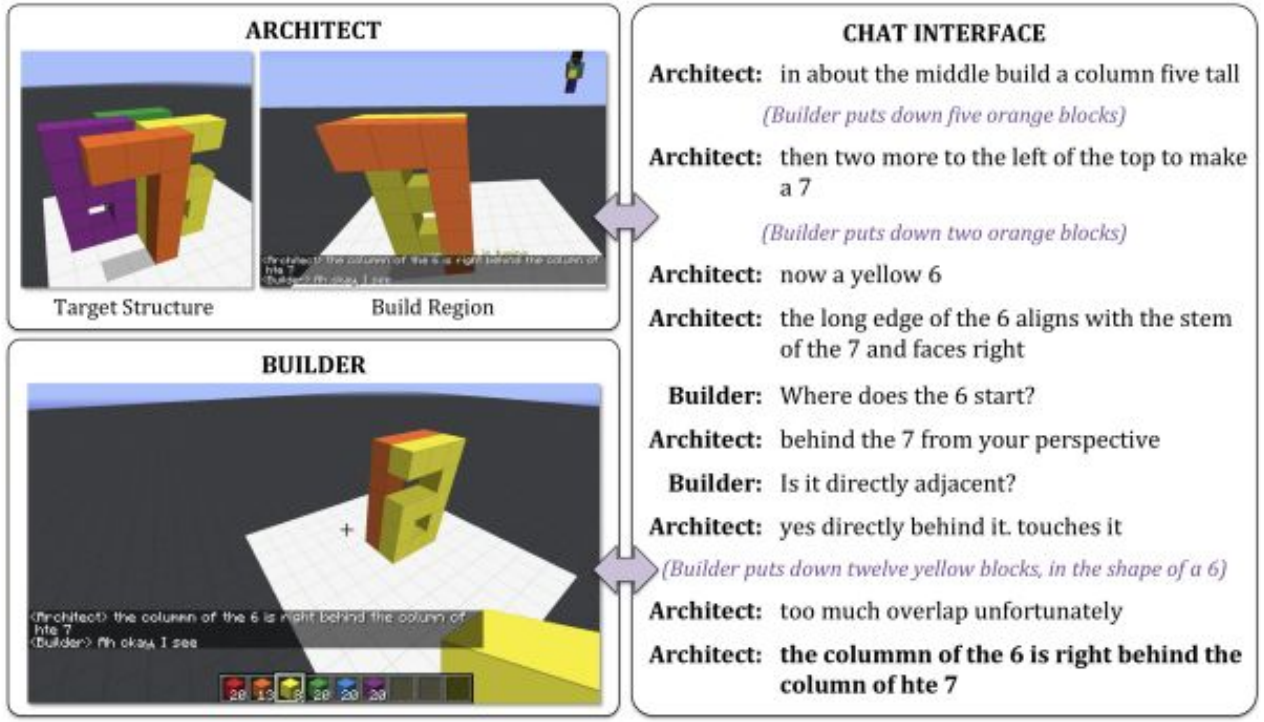
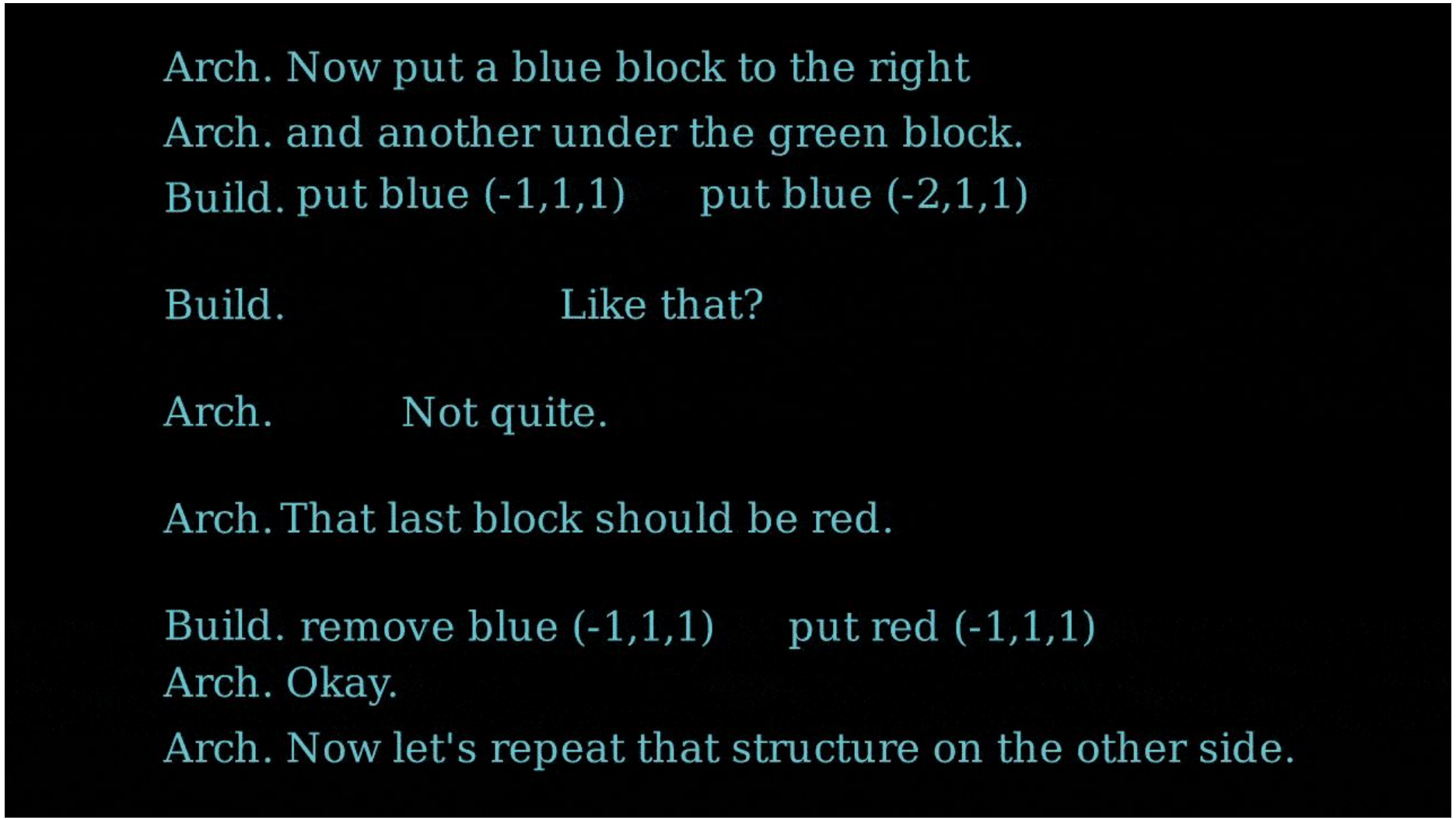


Figure 1: In the Minecraft Collaborative Building Task, the Architect (**A**) has to instruct a Builder (**B**) to build a target structure. **A** can observe **B**, but remains invisible to **B**. Both players communicate via a chat interface. (NB: We show **B**'s actions in the dialogue as a visual aid to the reader.)

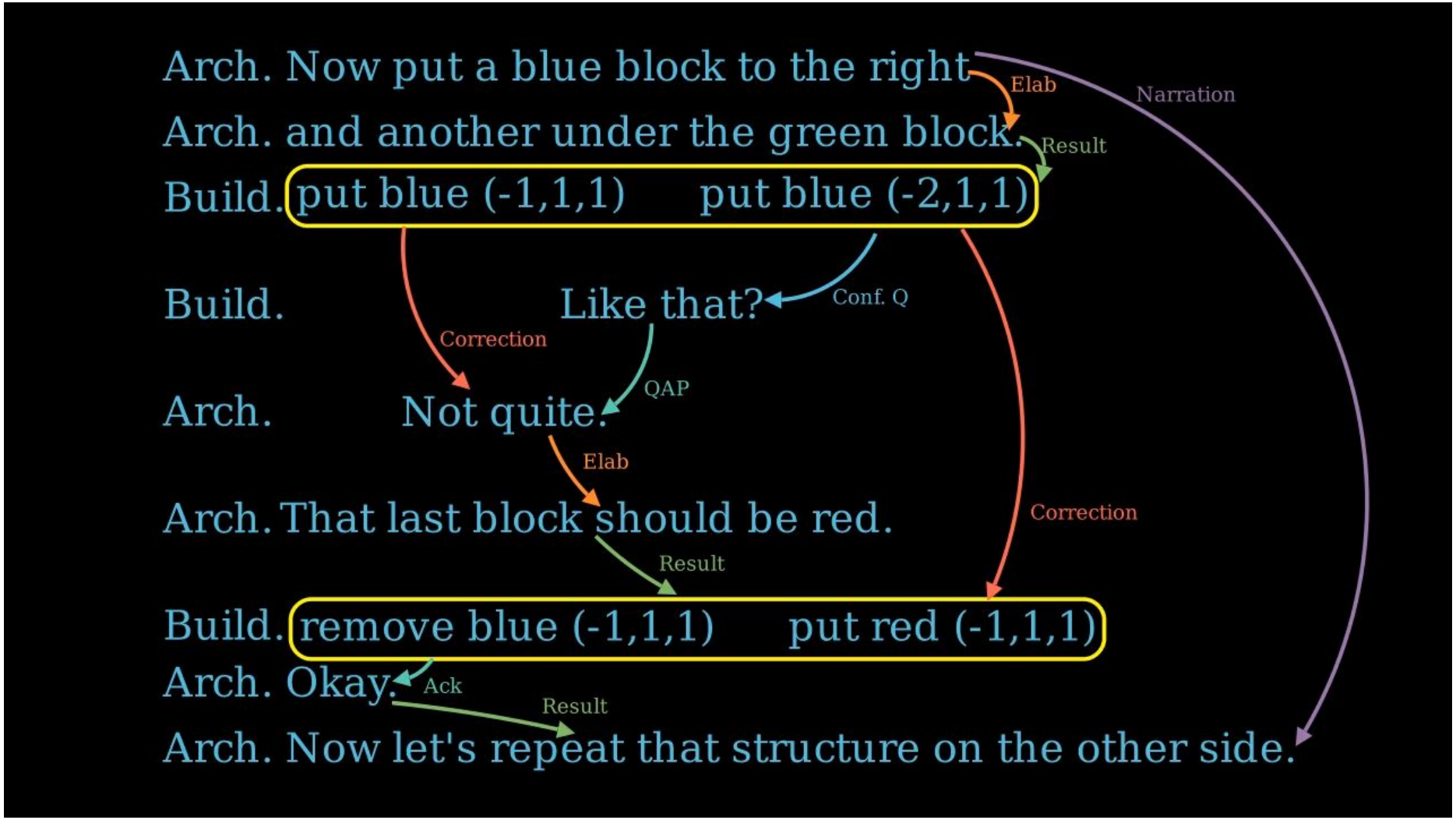
Anjali Narayan-Chen, Prashant, Jayannavar, and Julia Hockenmaier. 2019. Collaborative Dialogue in Minecraft. ACL.

The Minecraft Structured Dialogue Corpus



Thompson K., Hunter J., Asher N., Discourse Structure for the Minecraft Corpus. LREC, 2024

The Minecraft Structured Dialogue Corpus

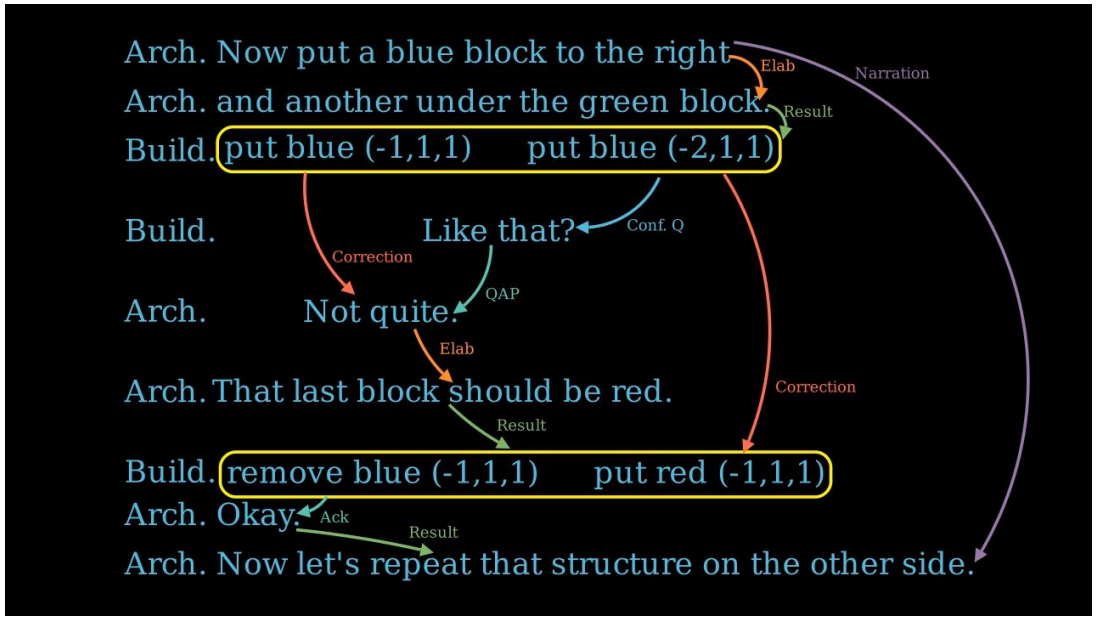


Thompson K., Hunter J., Asher N., Discourse Structure for the Minecraft Corpus. LREC, 2024

Discourse graphs

A discourse structure is a graph (V, E, l) with :

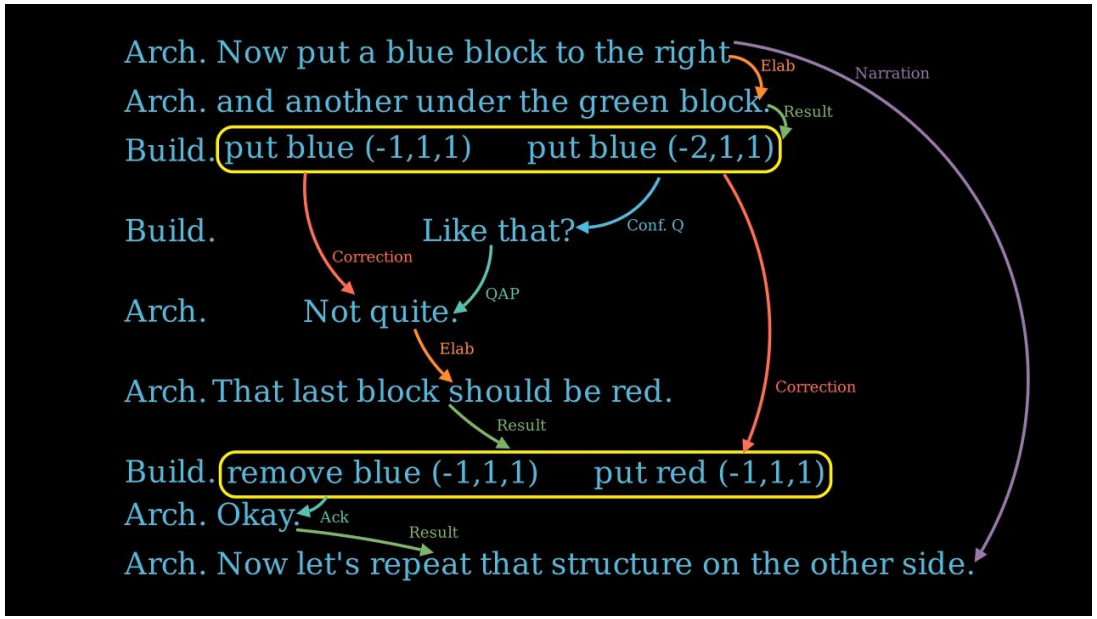
-A set of V discourse units $\{e_0, e_1, \ldots e_n\}$
-A set of edges $E \subset V \times V$
-l a labeling function $l : (e_k, e_i) \rightarrow r$
where $r \in R$
and R is a set of discourse relation types



Discourse parsing

A discourse structure is a graph (V, E, l) with :

-A set of V discourse units $\{e_0, e_1, \ldots e_n\}$
-A set of edges $E \subset V \times V$
-l a labeling function $l : (e_k, e_i) \rightarrow r$
where $r \in R$
and R is a set of discourse relation types

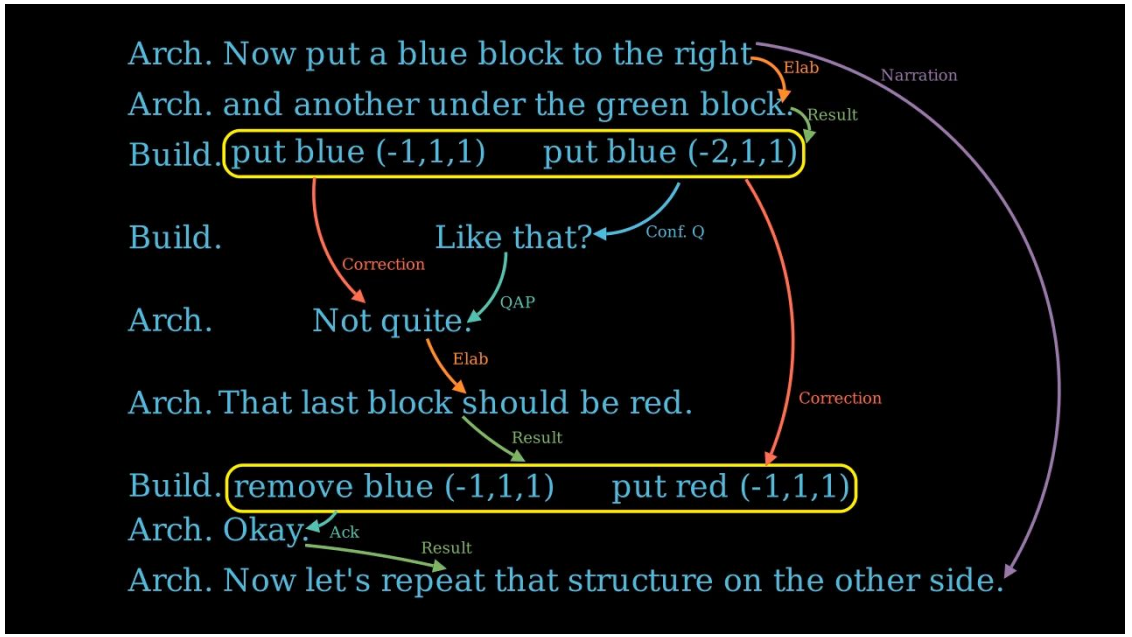


A discourse parser takes a dialogue D and outputs relation triples S.

$$S = \{ (e_k, e_i, r_{ki}), e_i \in V, \ e_i \in V, \ r_{ki} \in R \}$$

Discourse parsing

Local parsing looks at pairs of units in isolation:



Arch: Now put a blue block to the right

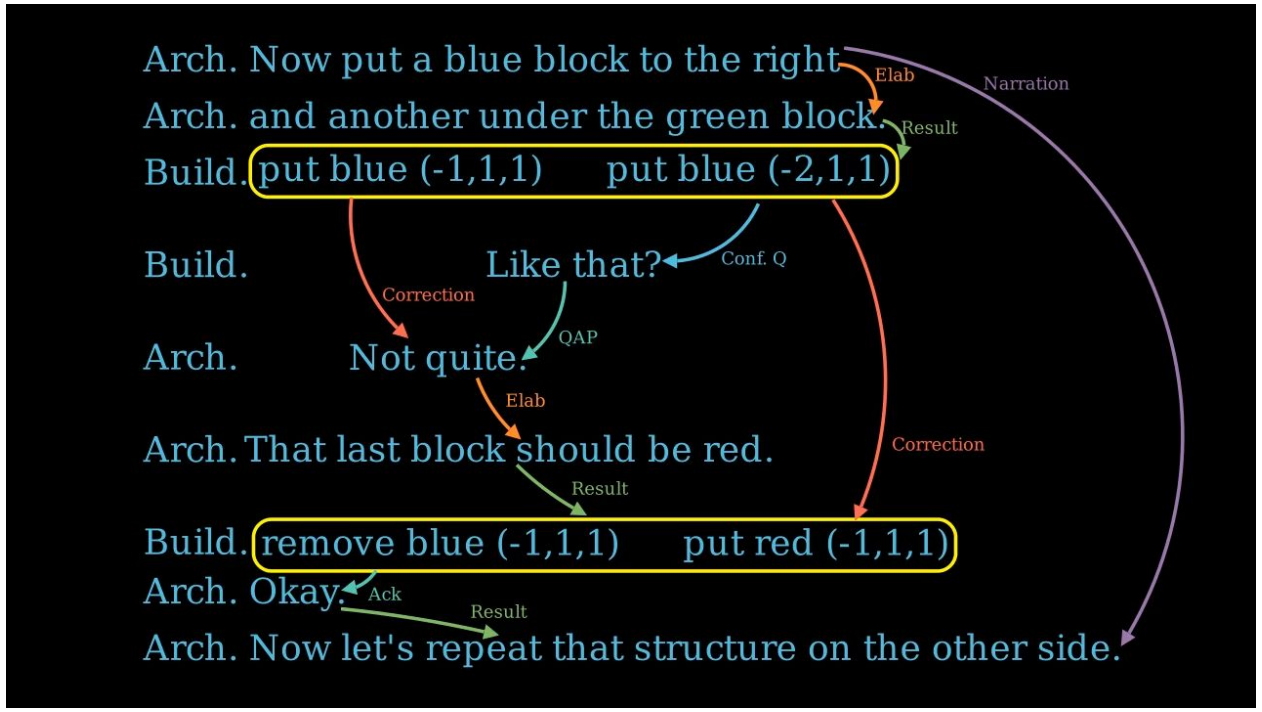
?→ Arch: and another under the green block
?→ Build: *put blue…put blue…*
?→ Build: Like that?
?→ Arch: That last block should be red.
?→ Build: *remove blue…put red…*
…

Discourse parsing

Arch: Now put a blue block to the right

?→ Arch: and another under the green block
?→ Build: *put blue…put blue…*
?→ Build: Like that?
?→ Arch: That last block should be red.
?→ Build: *remove blue…put red…*
?→ Arch: Okay.

Narration→ Arch: Now let's repeat that structure on the other side.



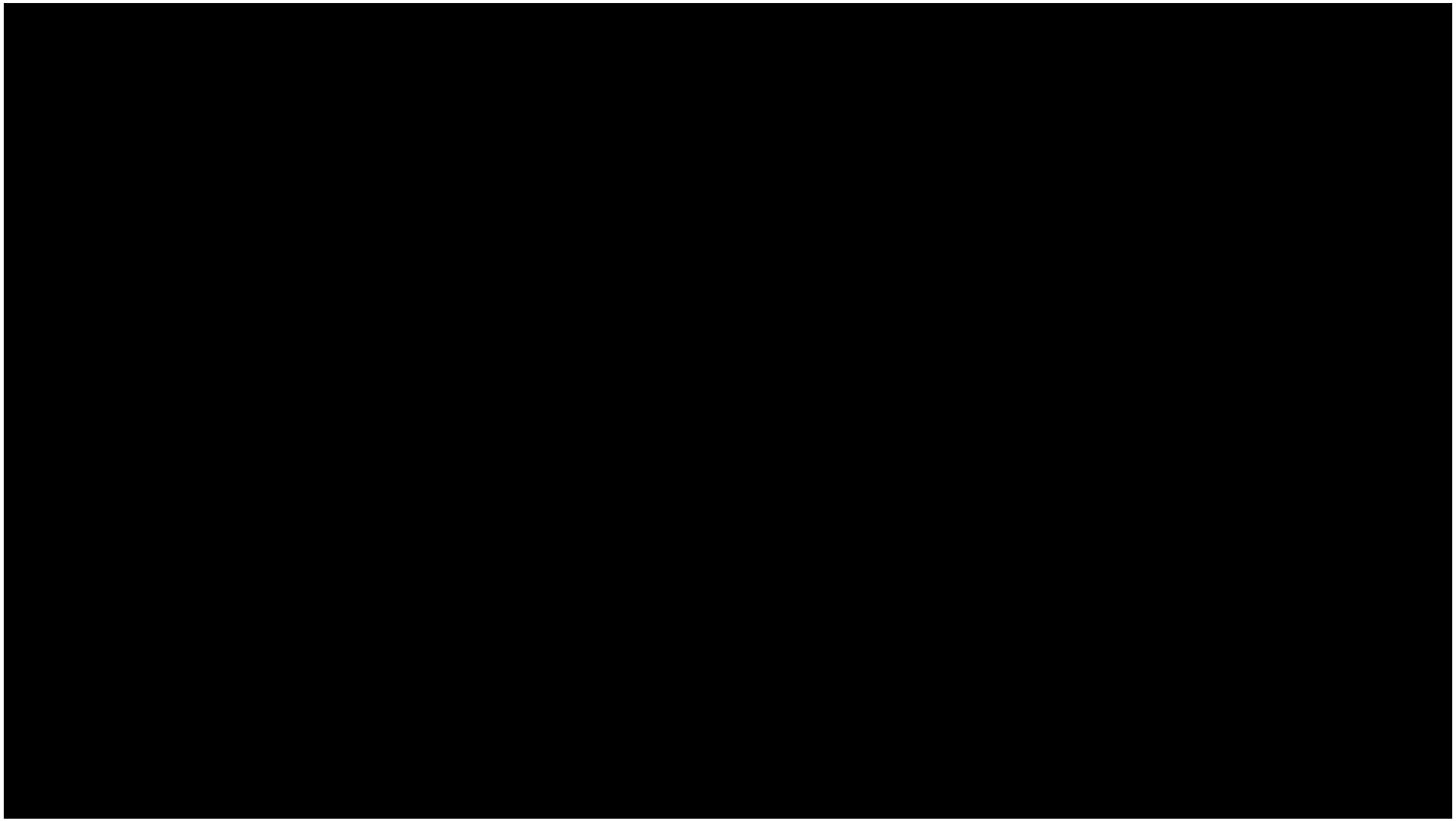
Discourse parsing as a generative task

To parse using an autoregressive model, we translate (D, S) into a pair of text sequences (x, y) where S is “linearized”, or represented as a sequence of triples, *e.g.* : ACK(0,1) CLARIFQ(0,2)

Let x denote the dialogue D and structure S until the current speaker turn. The model computes the conditional probability $p(y \mid x)$ where y is the structure between the new turn and the dialogue D.

Li et al. 2024. Dialogue Discourse Parsing as Generation: A Sequence-to-Sequence LLM Approach. SIGDIAL.

ANITI

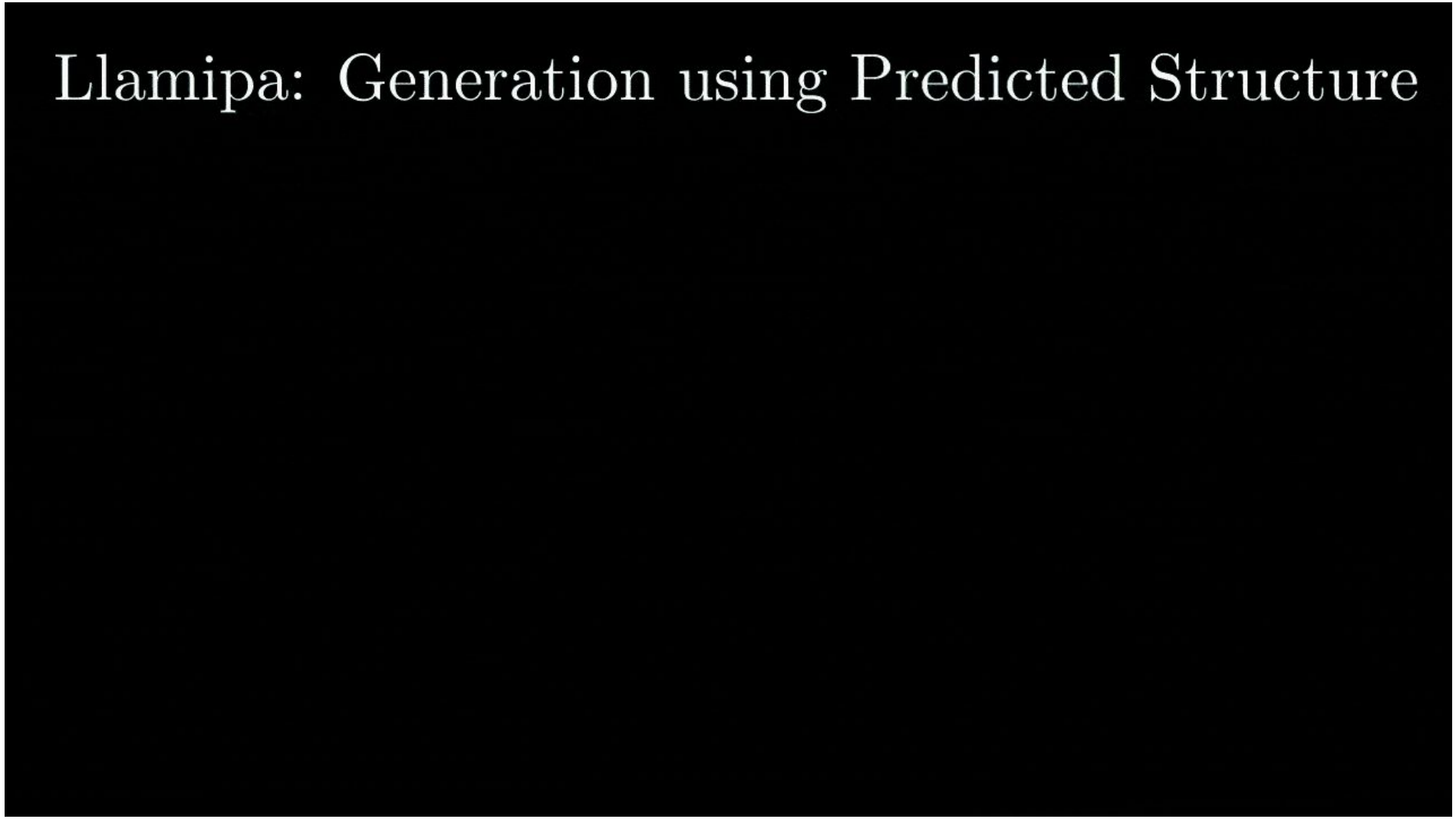


ANITI

Llamipa: Incremental Training

{Context: 0. Build: Mission has started.
1. Build: Hello
2. Build: What are we building today?
Structure: ACK(0,1) CLARIFQ(0,2) → Gold structure
New Turn: 3. Arch: so this looks like a table with tetris pieces on it}
###DS: QAP(2,3)

ANITI



ANITI

Llamipa: Generation using Predicted Structure

Identify the discourse structure (DS) for the new turn:

{Context: 0. Build: Mission has started.
Structure:
New Turn: 1. Build: Hello
2. Build: What are we building today?}
###DS: ACK(0,1) CLARIFQ(0,2)

{Context: 0. Build: Mission has started.
1. Build: Hello
2. Build: What are we building today?
Structure: ACK(0,1) CLARIFQ(0,2)
New Turn: 3. Arch: so this looks like a table with tetris pieces on it}
###DS: QAP(2,3)

ANITI

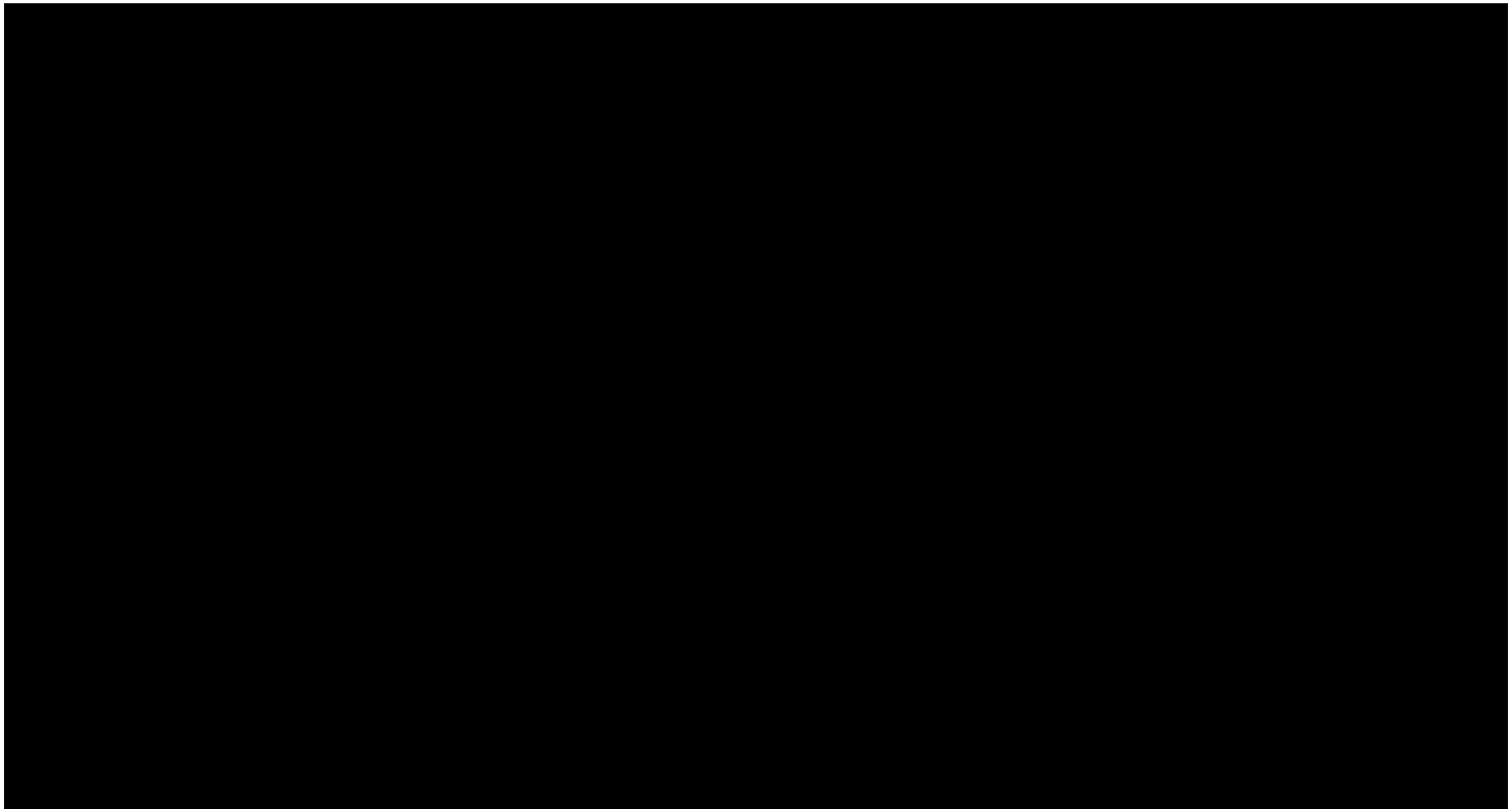
	MSDC		STAC Situated	
	Link	Link+Rel	Link	Link+Rel
Llamipa2+g	0.8561	0.7664	-	-
Llamipa2+p	0.8579	0.7570	-	-
Llamipa3+g	0.9004	0.8154	-	-
Llamipa3+p	**0.8830**	**0.7951**	**0.8612**	**0.7796**
BERTLine	0.7870	0.6901	0.7667	0.6788

ANITI

Is Llamipa using the explicit structure given in its context window?

Or is its performance due to the superior Llama embeddings?

Ablations



Ablations

Llamipa3+p	{Context: 0. Build: Mission has started. 1. Build: Hello 2. Build: What are we building today? Structure: ACK(0,1) CLARIFQ(0,2) New Turn: 3. Arch: so this looks like a table with tetris pieces on it} ###DS:
Llamipa+rand	{Context: 0. Build: Mission has started. 1. Build: Hello 2. Build: What are we building today? Structure: QAP(0,1) Elab(0,2) New Turn: 3. Arch: so this looks like a table with tetris pieces on it} ###DS:
Llamipa+empty	{Context: 0. Build: Mission has started. 1. Build: Hello 2. Build: What are we building today? Structure: New Turn: 3. Arch: so this looks like a table with tetris pieces on it} ###DS:
Llama3-local	{Context: 2. Build: What are we building today? New Turn: 3. Arch: so this looks like a table with tetris pieces on it} ###DS:

Ablations

	BERTLine	Llama3-local	Llamipa3+p	Llamipa3+g	Llamipa+rand	Llamipa+emp
Result	0.85	0.86	0.90	0.91	0.83	0.72
Acknowledgement	0.81	0.85	0.86	0.86	0.81	0.72
Narration	**0.50**	**0.54**	**0.82**	**0.91**	**0.28**	**0.18**
Elaboration	0.75	0.77	0.77	0.77	0.75	0.68
Correction	**0.31**	**0.64**	**0.73**	**0.80**	**0.49**	**0.52**
Continuation	0.44	0.47	0.49	0.50	0.44	0.33
Question-answer Pair	**0.76**	**0.80**	**0.82**	**0.83**	**0.42**	**0.14**
Comment	0.50	0.57	0.60	0.61	0.57	0.54
Confirmation-Question	0.86	0.89	0.93	0.93	0.91	0.89
Clarification-Question	0.61	0.66	0.73	0.73	0.70	0.41
Contrast	0.80	0.79	0.75	0.74	0.72	0.68
Question-Elaboration	0.39	0.36	0.30	0.36	0.33	0.33
Alternation	0.88	0.88	0.83	0.88	0.90	0.96
Explanation	0.00	0.00	0.00	0.00	0.00	0.00
Conditional	0.58	0.00	0.00	0.00	0.00	0.00
Sequence	0.00	0.00	0.00	0.00	0.00	0.05
Link+Rel F1	0.69	0.73	0.80	0.81	0.65	0.56
Link F1	0.78	0.82	0.88	0.90	0.77	0.72

Ablations

	BERTLine	Llama3-local	Llamipa3+p	Llamipa3+g	Llamipa+rand	Llamipa+emp
Result	0.85	0.86	0.90	0.91	0.83	0.72
Acknowledgement	0.81	0.85	0.86	0.86	0.81	0.72
Narration	**0.50**	**0.54**	**0.82**	**0.91**	**0.28**	**0.18**
Elaboration	0.75	0.77	0.77	0.77	0.75	0.68
Correction	**0.31**	**0.64**	**0.73**	**0.80**	**0.49**	**0.52**
Continuation	0.44	0.47	0.49	0.50	0.44	0.33
Question-answer Pair	**0.76**	**0.80**	**0.82**	**0.83**	**0.42**	**0.14**
Comment	0.50	0.57	0.60	0.61	0.57	0.54
Confirmation-Question	0.86	0.89	0.93	0.93	0.91	0.89
Clarification-Question	0.61	0.66	0.73	0.73	0.70	0.41
Contrast	0.80	0.79	0.75	0.74	0.72	0.68
Question-Elaboration	0.39	0.36	0.30	0.36	0.33	0.33
Alternation	0.88	0.88	0.83	0.88	0.90	0.96
Explanation	0.00	0.00	0.00	0.00	0.00	0.00
Conditional	0.58	0.00	0.00	0.00	0.00	0.00
Sequence	0.00	0.00	0.00	0.00	0.00	0.05
Link+Rel F1	0.69	0.73	0.80	0.81	0.65	0.56
Link F1	0.78	0.82	0.88	0.90	0.77	0.72

Ablations

	BERTLine	Llama3-local	Llamipa3+p	Llamipa3+g	Llamipa+rand	Llamipa+emp
Result	0.85	0.86	0.90	0.91	0.83	0.72
Acknowledgement	0.81	0.85	0.86	0.86	0.81	0.72
Narration	**0.50**	**0.54**	**0.82**	**0.91**	**0.28**	**0.18**
Elaboration	0.75	0.77	0.77	0.77	0.75	0.68
Correction	**0.31**	**0.64**	**0.73**	**0.80**	**0.49**	**0.52**
Continuation	0.44	0.47	0.49	0.50	0.44	0.33
Question-answer Pair	**0.76**	**0.80**	**0.82**	**0.83**	**0.42**	**0.14**
Comment	0.50	0.57	0.60	0.61	0.57	0.54
Confirmation-Question	0.86	0.89	0.93	0.93	0.91	0.89
Clarification-Question	0.61	0.66	0.73	0.73	0.70	0.41
Contrast	0.80	0.79	0.75	0.74	0.72	0.68
Question-Elaboration	0.39	0.36	0.30	0.36	0.33	0.33
Alternation	0.88	0.88	0.83	0.88	0.90	0.96
Explanation	0.00	0.00	0.00	0.00	0.00	0.00
Conditional	0.58	0.00	0.00	0.00	0.00	0.00
Sequence	0.00	0.00	0.00	0.00	0.00	0.05
Link+Rel F1	0.69	0.73	0.80	0.81	0.65	0.56
Link F1	0.78	0.82	0.88	0.90	0.77	0.72

Ablations

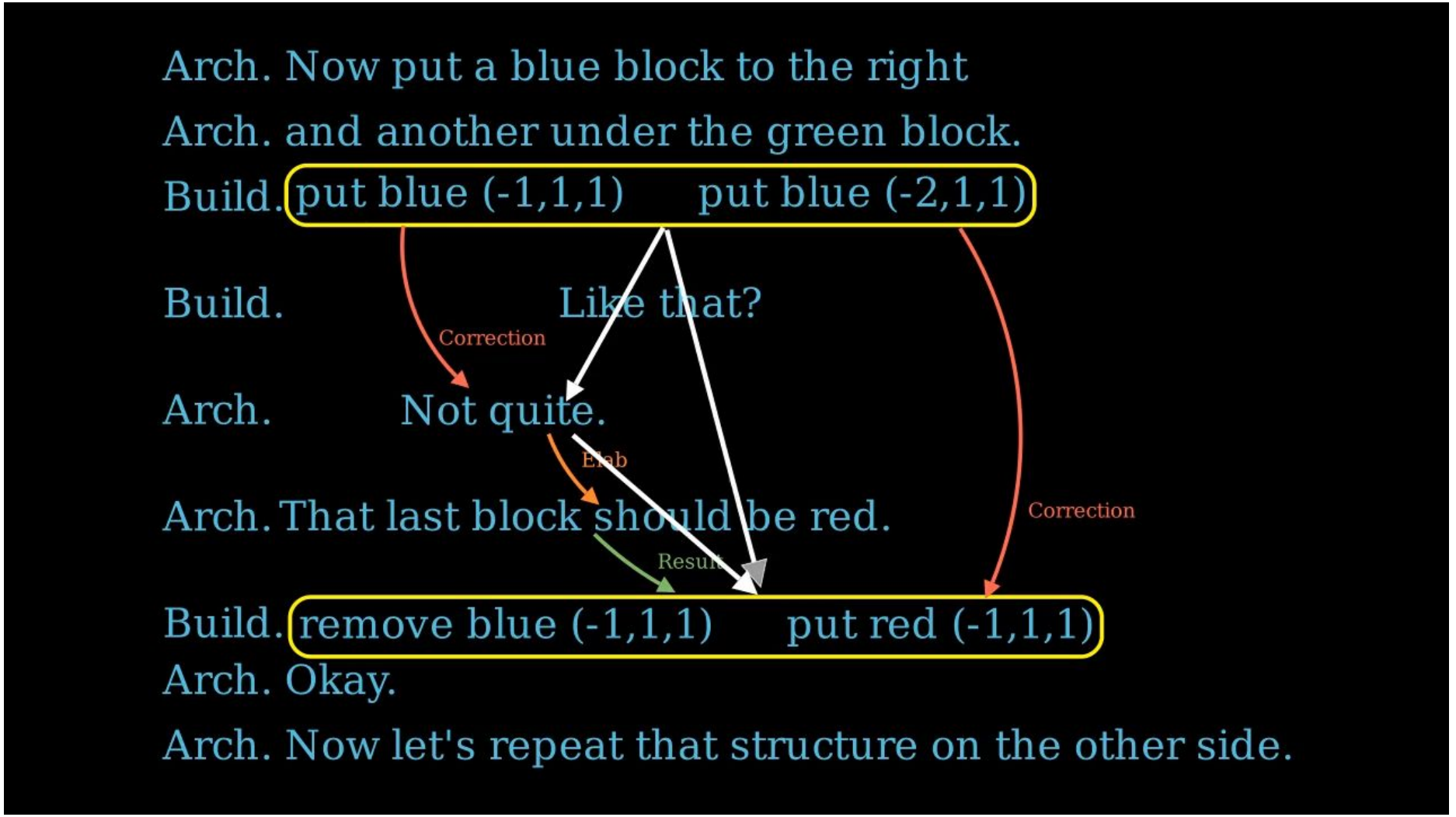
	BERTLine	Llama3-local	Llamipa3+p	Llamipa3+g	Llamipa+rand	Llamipa+emp
Result	0.85	0.86	0.90	0.91	0.83	0.72
Acknowledgement	0.81	0.85	0.86	0.86	0.81	0.72
Narration	**0.50**	**0.54**	**0.82**	**0.91**	**0.28**	**0.18**
Elaboration	0.75	0.77	0.77	0.77	0.75	0.68
Correction	**0.31**	**0.64**	**0.73**	**0.80**	**0.49**	**0.52**
Continuation	0.44	0.47	0.49	0.50	0.44	0.33
Question-answer Pair	**0.76**	**0.80**	**0.82**	**0.83**	**0.42**	**0.14**
Comment	0.50	0.57	0.60	0.61	0.57	0.54
Confirmation-Question	0.86	0.89	0.93	0.93	0.91	0.89
Clarification-Question	0.61	0.66	0.73	0.73	0.70	0.41
Contrast	0.80	0.79	0.75	0.74	0.72	0.68
Question-Elaboration	0.39	0.36	0.30	0.36	0.33	0.33
Alternation	0.88	0.88	0.83	0.88	0.90	0.96
Explanation	0.00	0.00	0.00	0.00	0.00	0.00
Conditional	0.58	0.00	0.00	0.00	0.00	0.00
Sequence	0.00	0.00	0.00	0.00	0.00	0.05
Link+Rel F1	0.69	0.73	0.80	0.81	0.65	0.56
Link F1	0.78	0.82	0.88	0.90	0.77	0.72

Ablations

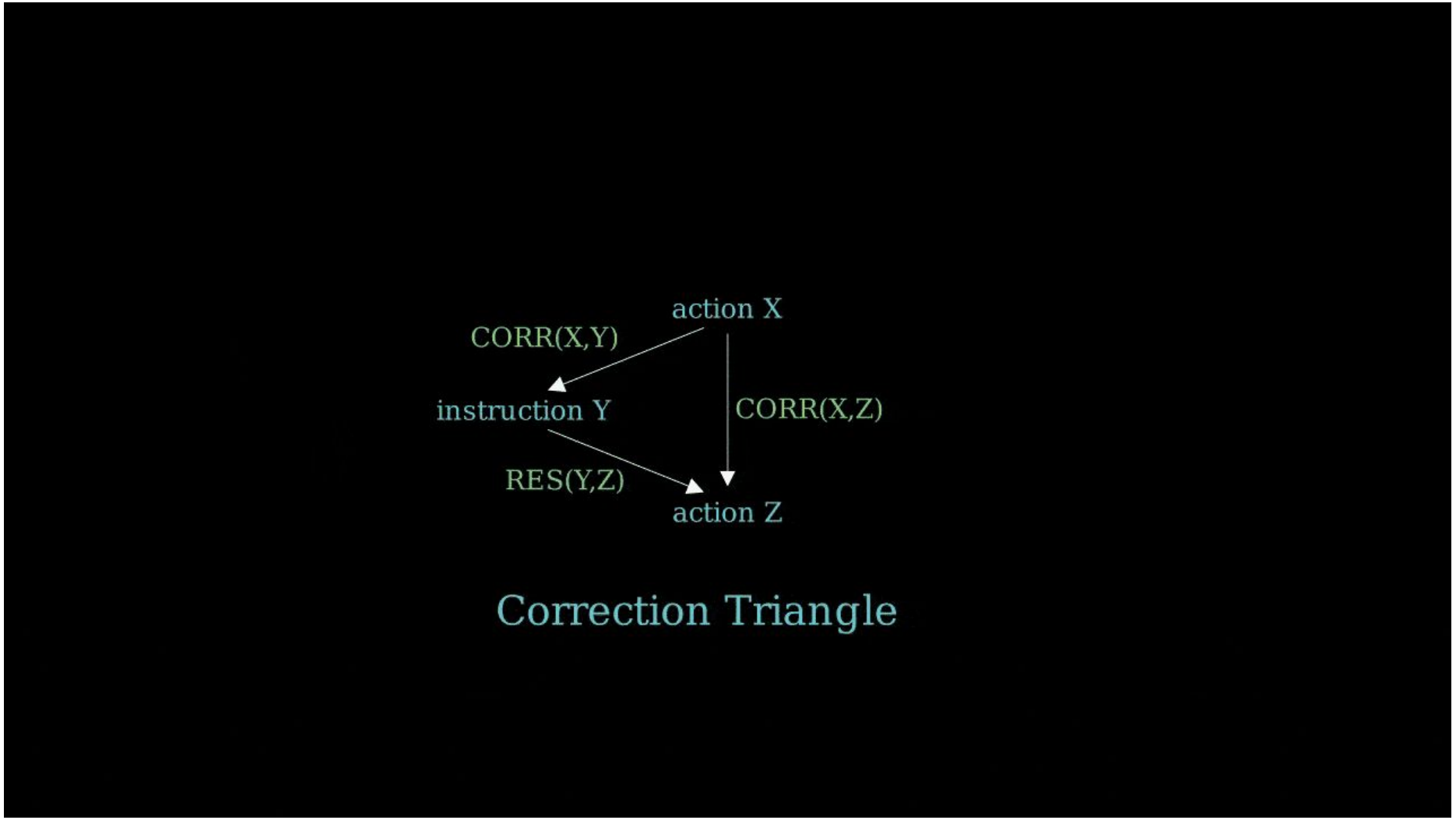
	BERTLine	Llama3-local	Llamipa3+p	Llamipa3+g	Llamipa+rand	Llamipa+emp
Result	0.85	0.86	0.90	0.91	0.83	0.72
Acknowledgement	0.81	0.85	0.86	0.86	0.81	0.72
Narration	**0.50**	**0.54**	**0.82**	**0.91**	**0.28**	**0.18**
Elaboration	0.75	0.77	0.77	0.77	0.75	0.68
Correction	**0.31**	**0.64**	**0.73**	**0.80**	**0.49**	**0.52**
Continuation	0.44	0.47	0.49	0.50	0.44	0.33
Question-answer Pair	**0.76**	**0.80**	**0.82**	**0.83**	**0.42**	**0.14**
Comment	0.50	0.57	0.60	0.61	0.57	0.54
Confirmation-Question	0.86	0.89	0.93	0.93	0.91	0.89
Clarification-Question	0.61	0.66	0.73	0.73	0.70	0.41
Contrast	0.80	0.79	0.75	0.74	0.72	0.68
Question-Elaboration	0.39	0.36	0.30	0.36	0.33	0.33
Alternation	0.88	0.88	0.83	0.88	0.90	0.96
Explanation	0.00	0.00	0.00	0.00	0.00	0.00
Conditional	0.58	0.00	0.00	0.00	0.00	0.00
Sequence	0.00	0.00	0.00	0.00	0.00	0.05
Link+Rel F1	0.69	0.73	0.80	0.81	0.65	0.56
Link F1	0.78	0.82	0.88	0.90	0.77	0.72

Correction Triangles

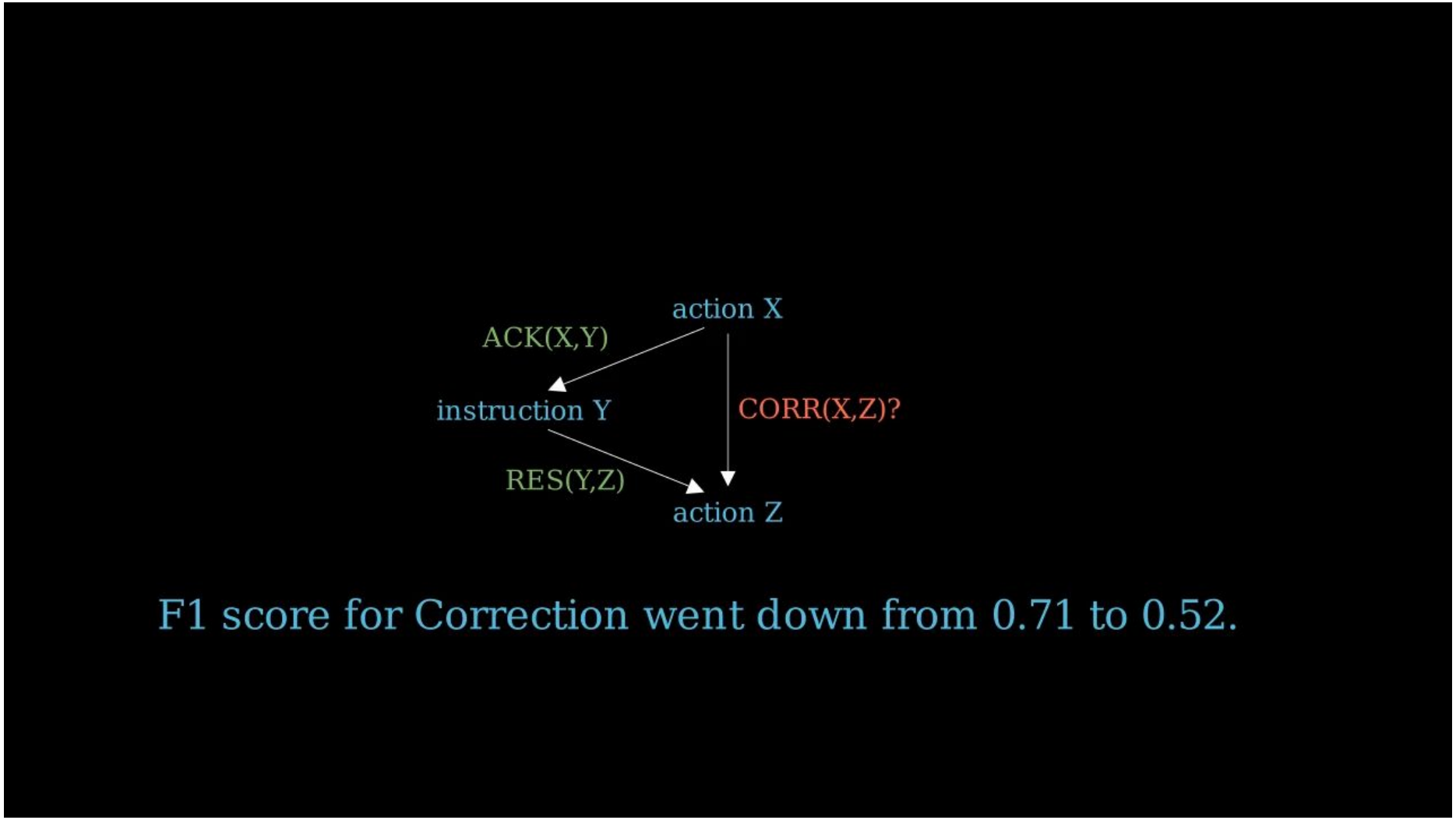
Correction Triangles



Correction Triangle Ablation



Correction Triangle Ablation



ANITI

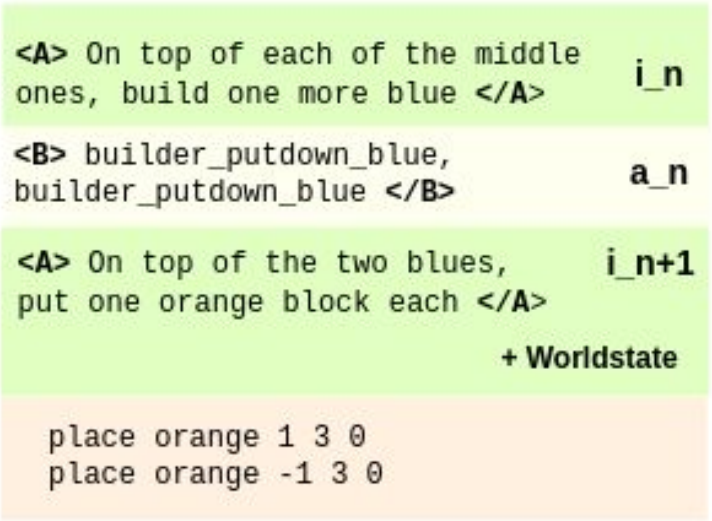
Generative approach to the action prediction task

Action Prediction Task

Predict the builder action sequence, a_{n+1}, given the conversation history.

Neural Builder is the **baseline** model for the MDC. It inputs a 3D CNN encoding of worldstate (blocks already on the grid) and two previous Architect instructions and intervening Builder moves to a GRU to predict a_{n+1}.

It gets a net-action F1 of 0.20.



Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute instructions in a Minecraft dialogue. ACL.

ANITI Nebula

But $i_n a_n i_{n+1}$ is suboptimal. It doesn't encode enough dialogue context.

LLMs allow us to give the entire conversation history as input.

We finetuned 3 versions of Llama using QLoRA on this task by giving *the entire dialogue history* as input. The best performing model, Llama-3-8b, almost doubles the net-action F1 score from 0.20 to 0.39. We refer to this model as Nebula.

Dataset	Llama-2-7b	Llama-2-13b	Llama-3-8b
Validation	0.292	0.323	0.398
Test	0.326	0.338	0.392

Issues with net-action F1

For each predicted output sequence, net-action F1 counts the new blocks placed on the grid as **correct *only*** if they match **the *exact* positions** of the blocks in the ground truth sequences.

This exact match criterion is far too rigid if we consider that natural language instructions are typically underspecified, e.g., *place a blue block in a corner.*

Synthetic data

Two levels:

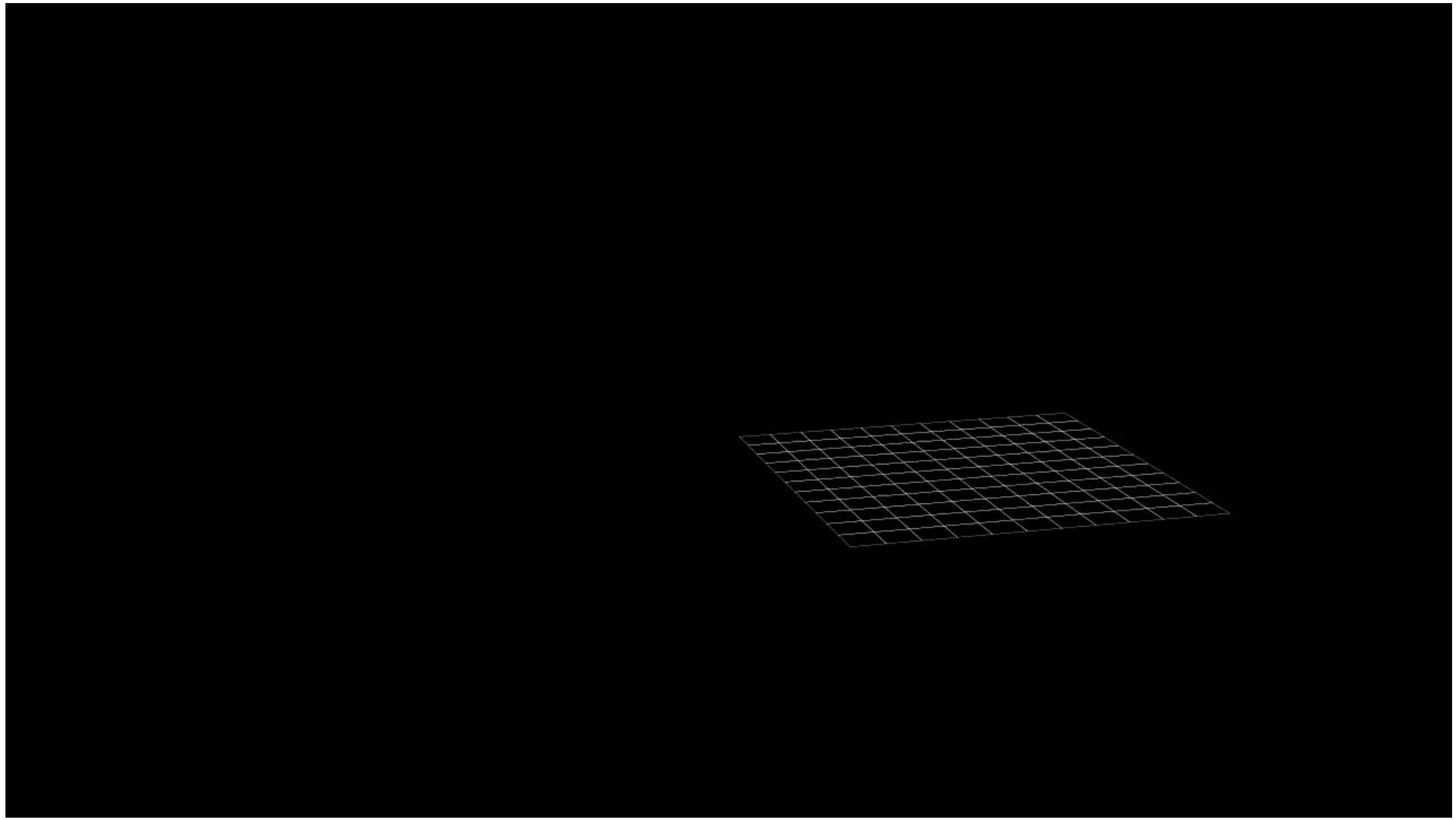
1. Level 1: Basic shapes such as rows, towers, rectangles, squares etc.

2. Level 2: Given a level-1 shape, place/remove a block in relation to it (e.g. place on top of, place to the side of, remove the bottom block, remove the middle block).

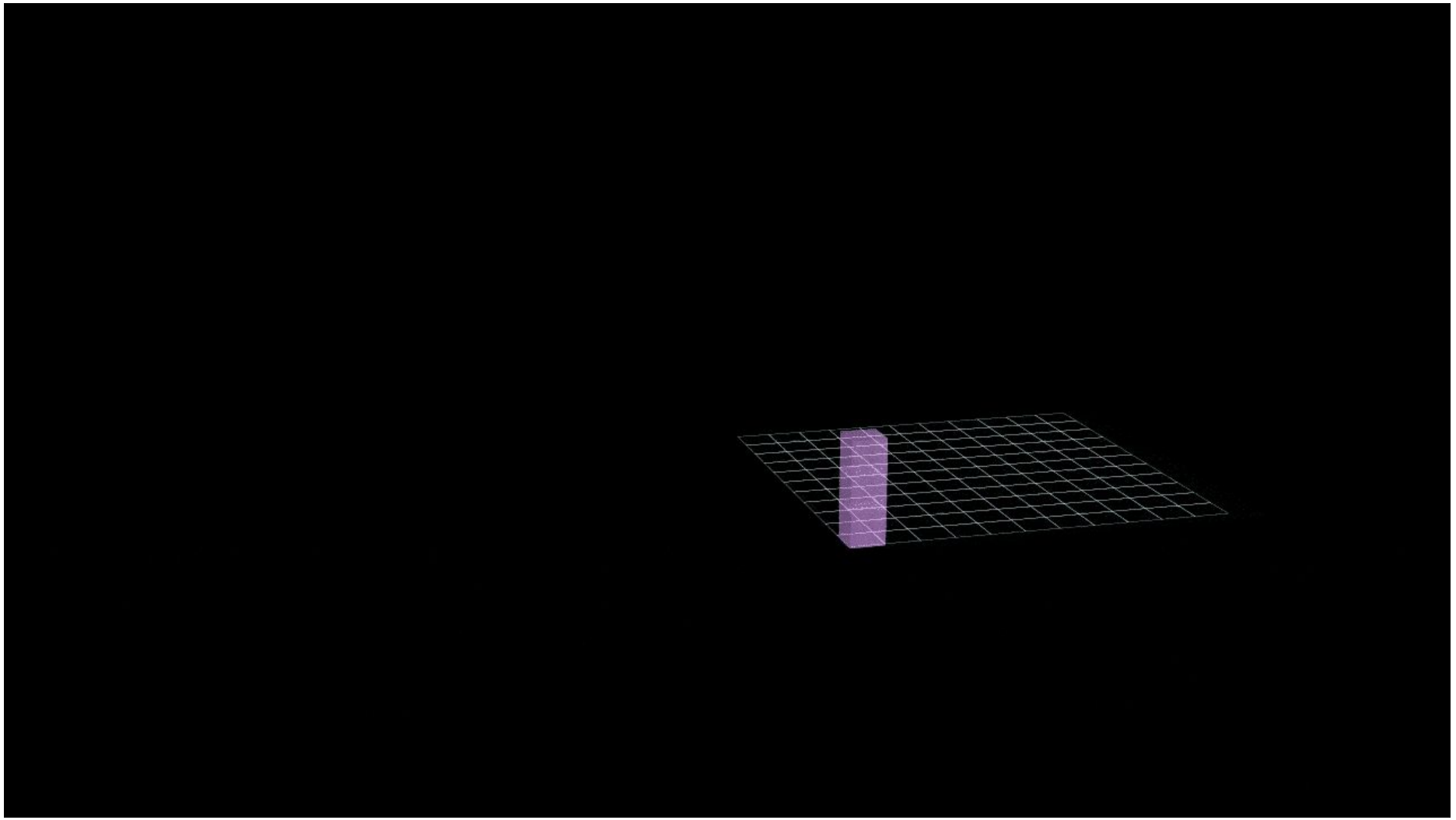
For both the levels, we evaluate the model performance by making use of **binary functions** like is_square(), is_ontopof() etc, thereby allowing for **multiple correct instantiations** for a given instruction.

ANITI

Level-1



ANITI Level-2



ANITI

Integrating discourse for action prediction

Is the entire conversation history needed?

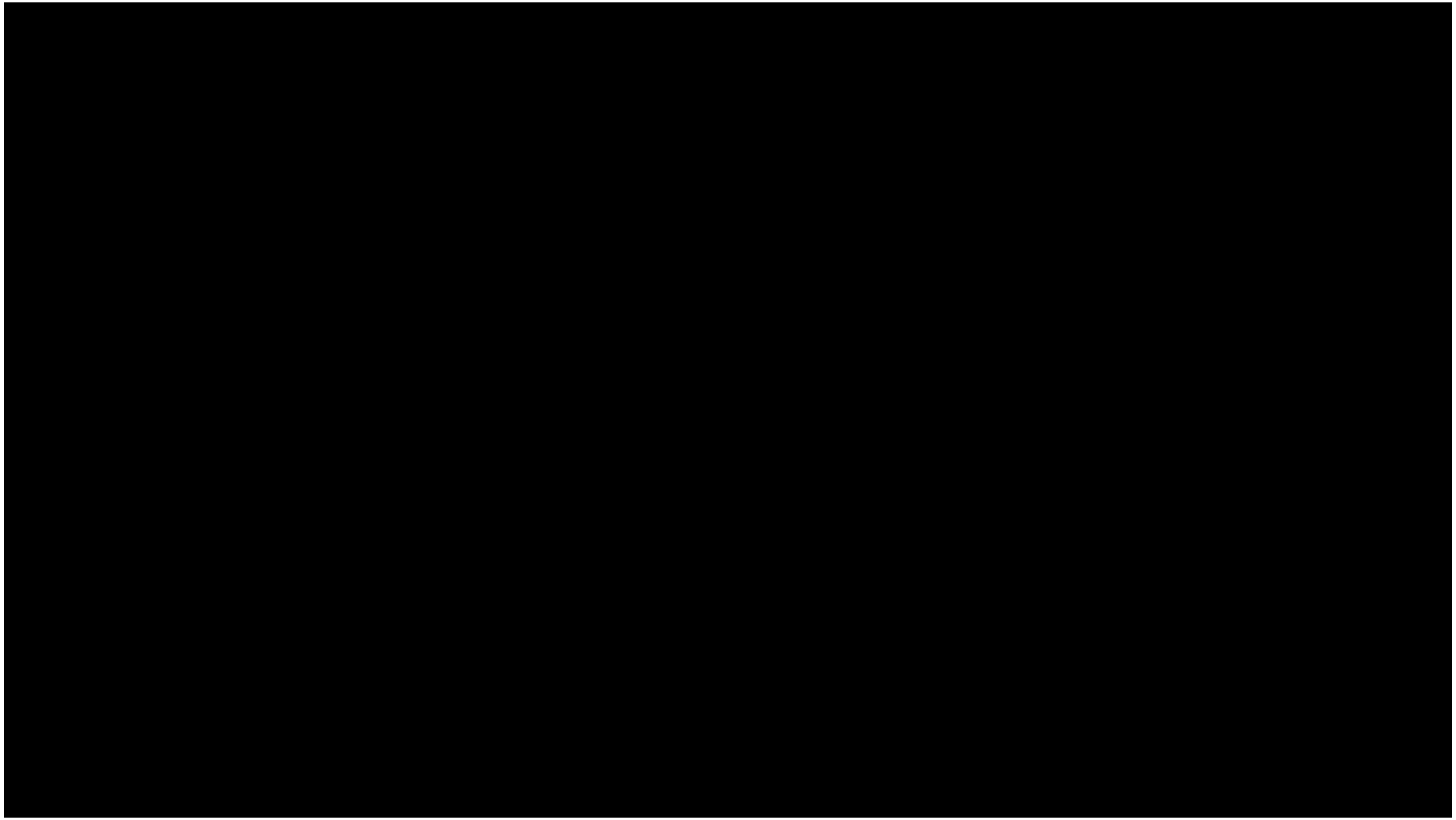
An alternative is to use narrative chunks from MSDC corpus as input to the model.

Narration relation connect two high-level instructions.

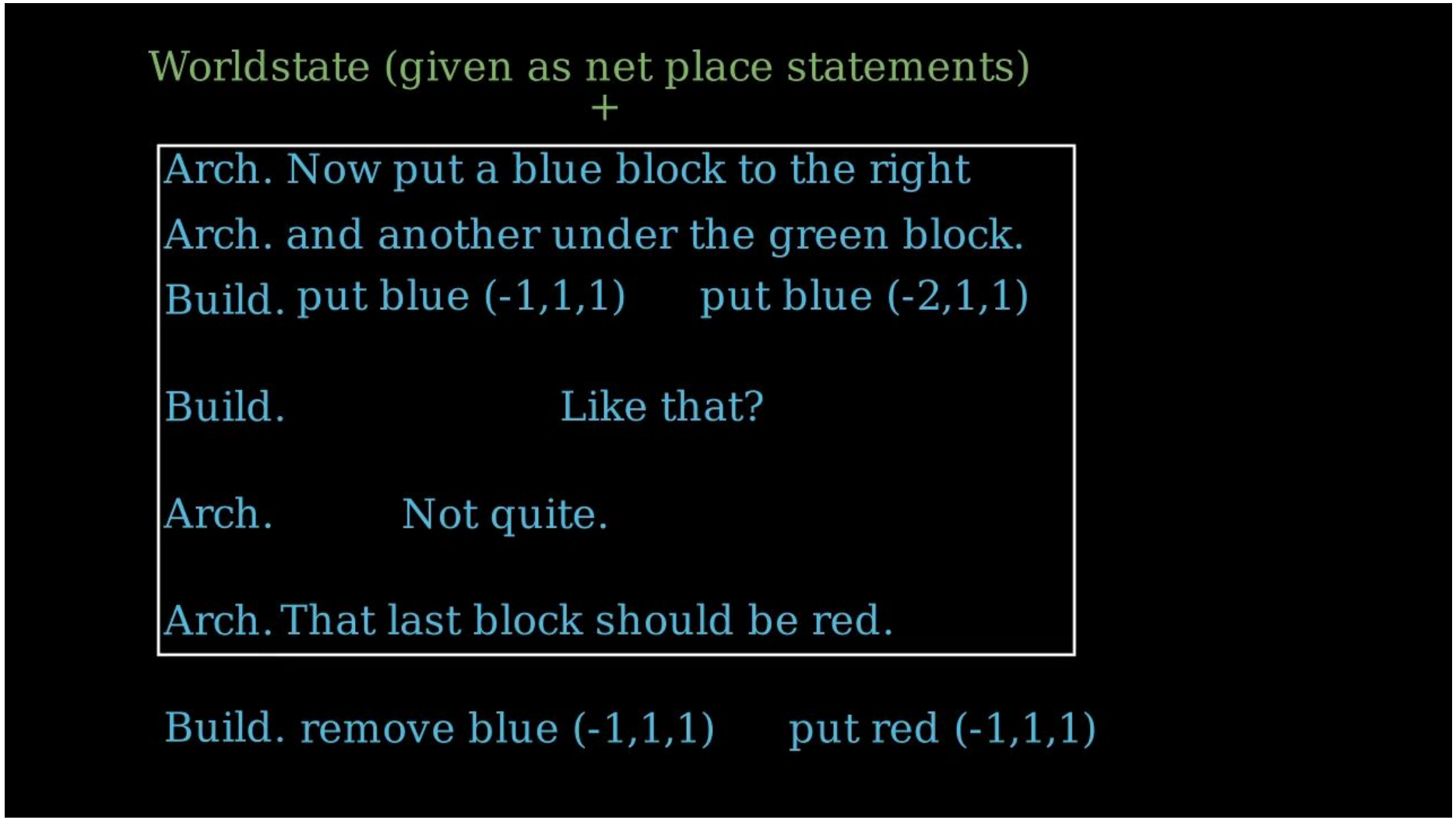
The dialogue chunk within a narrative arc is likely to have all the relevant information for the action prediction task.

Thompson K., Hunter J., Asher N., Discourse Structure for the Minecraft Corpus. LREC, 2024

Narrative arcs



Narrative arcs



Narrative arcs results

Model	Net-Action F1
Nebula	0.392
Nebula + N	0.380
Nebula + N/N	0.349
Nebula + $N/i_n a_n i_{n+1}$	0.311

The last two scores (0.349 and 0.311) are for cases where $N>i_n a_n i_{n+1}$

ANITI

Nebula

Llamipa

Merci!

IRIT
Institut de Recherche
en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

anr

LINAGORA