Collaborative Dialogue in Minecraft: Supplementary Material

Anjali Narayan-Chen[*] **Prashant Jayannavar**[*] **Julia Hockenmaier**
University of Illinois at Urbana-Champaign
{nrynchn2, paj3, juliahmr}@illinois.edu

1 Introduction

This document contains the supplementary materials accompanying the main paper, Collaborative Dialogue in Minecraft. First, we provide two example human-human dialogues from the Minecraft Dialogue Corpus, which is described in detail in Section 4 of the main paper, in Section 2. Section 3 details all hyperparameters of the models reported in Section 9 of the main paper. In section 4, we provide a complete description and results of the human evaluation described in Section 9.2 of the main paper. Section 5 provides a more in-depth qualitative analysis of two examples of generated utterances from our best model to accompany Section 9.3 of the main paper. Finally, term-specific word and synonym lists can be found in the Appendix.

2 Example Dialogues from the Corpus

Below, we include two sample dialogues from the corpus along with corresponding screenshots of the target structure to be built. Builder actions are interleaved in the dialogue in bolded italics.

Example 1.

A: hi
B: Hello
A: ready to go???
B: Yes I am!
A: this is either a chicken or a gun turret
A: start with blue
B: Excellent
A: put a line of three down
B puts down a row of three blue blocks
A: now parallel to that with one blank space add another line
B: Same color?
A: yep
B puts down a row of three blue blocks
A: now put a single block on each middle one
B puts down two blue blocks

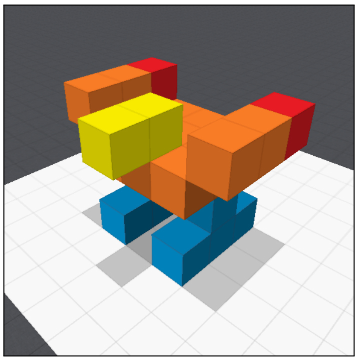


Figure 1: Target structure for Example 1.

A: good. that's the feet/legs
A: now orange
B: Okay
A: a flat 3x3 square centered over the blue parts
B puts down a square of nine orange blocks
B: Like this?
A: great
A: now wings. two orange blocks in a row up and diagonal aaway from the body
B puts down two orange blocks, then removes the first
B: Is this the right place?
A: yep. now one more closer to you
B puts down three and removes two orange blocks
A: great now mirror it
B: Copy that
B builds a mirror of the existing wing
A: now a single red block behind each of the wings. still in line with the oranges
B puts down two red blocks
A: then a row of two yellow blocks on the orange side of the wings for the neck/head
A: juts out over the edge
A: parallel to the wings
B: Should they be right next to the wings?
A: centered. one block in the middle edge orange
A: then another block coming away from the body
B puts down two yellow blocks
B: Here?
A: awesome
B: Cool
A: that's it

Example 2.

B: what is this
A: It looks weird, so I don't really know.

[*] Both authors equally contributed to the paper.

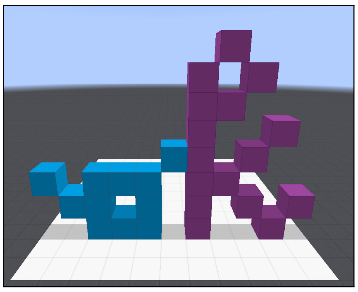


Figure 2: Target structure for Example 2.

B: ok

B: then guide me blok by block

A: Ok, starting on an edge, make a purple block two blocks in

B puts down a purple block, picks it up, and places it again

B: I'll stay with this perspective

B: unless you want me to change

A: Sorry, one block closer to the edge

B removes and places the purple block

A: Can you make a block 1 off the ground on the edge

B: like on top of that purple block?

A: Like a v with the purple block as the poiunt

B creates a v-shape using floating blocks

B: yeha?

A: Like that, except perpendicular to the edge

B begins to correct the orientation of the v, using a series of block placements and removals

B: there?

A: Yup, can you extend the block on the right diagonally up?

B continues the v to the right

B: like that

A: Yes, can you make a 7 block tall pillar to the right of that block. It should start from the ground

B: what color?

B: purple?

A: Purple still

B builds a pillar of 7 purple blocks

A: Can you extend the top block 2 blocks wtowards the edge

B: which edge

A: The side with the v

B puts down two purple blocks

A: Can you add a block on top and below the middle block the blocks we just placeD?

B puts down two purple blocks

B: um

A: Yes?

B: is this it?

A: Oh, yeah, can you punch out the middle block

B removes a purple block

A: Cool, from the block 3 blocks below that, can you extend it two blocks diagonally towards the closest edge?

B: what

B: literally what

A: Sorry,

A: You know how we have the block next to the pillar?

A: Close to the bottom

B: yes

B: the one i am looking at?

A: Can you add a block to it going diagonally.

B: down?

B builds a floating purple block

B: like that

A: One beloow the block you just placed, sorry.

B puts down a purple block, then removes the previous purple block

A: And another block to the left and up

B builds a floating purple block

A: Ok, that is all of the purple blocks. We are going to use blue blocks now.

A: to the right of the pillar,

B: is there symmetry?

A: No,

B: sad

A: I know, maybe it will make sense to you. There is a block attached the the pillar on the fourth block from the ground

B puts down a blue block

B: there

A: Perfect

A: From the right of that, can you make a 3 block pillar from the ground

B builds a pillar of three blue blocks

B: is this a 2D structure?

A: Yes

A: Can you make a ring using the pillar we just made?

B: 3x3

B: ?

A: Yes

A: sorry,

B puts down five blue blocks

B: that

A: Yup, on the middle block of the ring's right side, can you put a blue block?

B puts down a blue block

A: And up and to the right of that, one more blue block

B builds a floating blue block

B: it looks like a cursive "ok" form the other end

A: That should be it. Oh, I never learned cursive.

3 Model Hyperparameters

We use Gated Recurrent Units (GRUs) ((Chung et al., 2014)) for all RNN modules and use 300-dimensional pretrained GloVe word embeddings ((Pennington et al., 2014)). All linear layers were initialized using Xavier initialization ((Glorot and Bengio, 2010)). All RNNs have a hidden state size of 300. In cases where we use a bidirectional encoder RNN, the sum of the two final hidden states in either direction constitutes the final encoding of dialogue history (used to initialize the decoder).

Ablation study For the ablation study, we analyze the effect of adding various block counters representations to a fixed dialogue history representation. Thus, the RNN framework we use for all models in the ablation study is a 2-layer bidirectional encoder RNN connected to a single-layer decoder RNN. All models were trained with dropout of 0.5 for both the RNNs and the counter embedding layers. Specific hyperparameter configurations for the models in Table 1 are as follows:

- **seq2seq** (row 1): The baseline sequence-to-sequence model conditioned only on full dialogue history features the abovementioned RNN framework.
- **+ local only** (row 2): The model incorporating only local block counters features a counter embedding size of 200.
- **+ global only** (row 3): The model incorporating only the global block counters features a a counter embedding size of 15.
- **+ global & local** (row 4): Our final model concatenates both local and global counters and embeds them into a 200-dimensional vector.

Test set We optimize the seq2seq and full models by performing a grid search over model hyperparameters. Specific hyperparameter configurations for the models in Table 2 are as follows:

- **seq2seq** (row 1): The baseline model features a single-layer bidirectional encoder RNN connected to a single-layer decoder RNN with a dropout of 0.5 for both RNN modules.
- **+ global & local** (row 2): Our final model features a counter embedding size of 200. It was trained using dropout=0.5 for the counter embedding layers and dropout=0 for RNN modules.

4 Human Evaluation

For each of the human evaluation criteria, we include here full descriptions of the evaluation guidelines as well as evaluation results and inter-rater reliability metrics of human judgments using Krippendorf's alpha (Hayes and Krippendorff, 2007).

4.1 Fluency

Evaluators were asked to rate the fluency of an utterance by selecting one of the following categories:

- *Perfectly fluent:* the utterance contains no spelling or grammatical mistakes and is well-formed in the context of English text chat. The utterance may not necessarily consist of complete sentences, but consists of long enough sentences to remain reasonably grammatical given the dialogue context.

Model	Yes	Somewhat	No
seq2seq	97.0	3.0	0.0
+ global & local	93.0	5.0	2.0
human	83.0	17.0	0.0

Table 1: Percentage of utterances deemed fluent by majority vote across 3 human evaluators.

- *Somewhat fluent/disfluent:* the utterance contains mistakes but still contains parts that resemble fluent English chat. Mistake types can include: 1) typos, 2) inappropriate use or addition of punctuation, 3) run-on sentences, unnecessary repetition, 4) inappropriately dropped words, etc.
- *Completely disfluent:* the utterance is word salad.

Fluency results are shown in Table 1 ($\alpha = 0.774$). While models are trained to produce mostly syntactically mistake-free utterances, humans are prone to producing utterances with typos and sentence fragments in a text chat.

4.2 Dialogue Act Annotation

Evaluators were asked to choose all dialogue acts from a predefined set that categorized any given utterance. The predefined categories, determined after a manual qualitative analysis of utterances in the validation set, are as follows:

- *Instruct **B**:* the utterance instructs **B** to move, place or remove blocks, or otherwise execute some action within the game using their player character (*"Place a red block"*, *"Move around to the left corner"*) ($\alpha = 0.884$)
- *Describe Target:* the utterance provides a description of the target structure or elements of it, such as blocks and substructures within the target (*"We're going to build a 3x3"*, *"Next we'll do wings"*) ($\alpha = 0.713$)
- *Answer question:* the utterance provides a response to a question posed by **B** (*"3 high"* in response to *"How tall?"*, *"perfect!"* in response to *"Is this right?"*) ($\alpha = 0.802$)

- *Confirm **B**'s actions or plans:* the utterance provides a confirmation (*"yes"*, *"that's right!"*, etc.) or rejection (*"no"*, *"sorry"*, etc.) in response to actions that **B** has taken or plans proposed/executed by **B** ($\alpha = 0.696$)

- *Correct or clarify **A** or **B**:*[1] the utterance rectifies mistakes made by **A** or **B** or provides additional clarifying information (*"No, get rid of the last block you placed"*, *"one more over to the left"*) ($\alpha = 0.778$)

- *Other:* other utterance types not covered by the above categories, including but not limited to: offhand comments, chitchat, greetings, etc. (*"Hello Builder"*, *"Haha, I couldn't see that side"*) ($\alpha = 0.804$)

Dialogue act annotation results can be found in Table 3 of the main paper. Associated analysis and discussion can be found in Section 9.2.

4.3 Appropriateness

Evaluators were asked to rate the appropriateness of an utterance by categorizing the appropriateness of the *type* of utterance in the game context into one of the following categories:

- *Appropriate:* the type of utterance is a completely reasonable response given the preceding dialogue; e.g., if a question was asked, the utterance answers it; if confirmation is requested, the utterance provides it; etc.

- *Maybe appropriate:* the type of utterance could be considered a reasonable response given the preceding dialogue; though it may not be the most natural or polite option, it is not clearly an incorrect type of response that should be elicited from the dialogue.

- *Inappropriate:* the type of utterance is clearly incorrect given the preceding dialogue.

- *N/A:* the utterance cannot be evaluated for appropriateness (due to disfluency).

Appropriateness results are shown in Table 2 ($\alpha = 0.588$). Because of the tendency for models to routinely generate instructions, model responses were seen as slightly inappropriate and dismissive of the dialogue context. On the other hand, human responses, containing a wider spread of dialogue act types, were almost universally seen to be appropriate in context.

Model	Yes	Maybe	No	N/A
seq2seq	87.0	11.0	0.0	2.0
+ global & local	87.0	12.0	0.0	1.0
human	97.0	2.0	0.0	1.0

Table 2: Percentage of utterances deemed appropriate by majority vote across 3 human evaluators.

		Somewhat	Unclear/
Model	Clear	Unclear	Impossible
seq2seq	63.6	29.9	6.5
+ global & local	61.6	30.1	8.2
human	91.7	8.3	0.0

Table 3: Percentage of instruction-type utterances deemed executable by majority vote across 3 human evaluators. Instruction-type utterances are identified by majority vote of annotated dialogue acts.

4.4 Executability

Evaluators were asked to rate the executability of instruction-type utterances in the current game state. This criterion aimed to analyze the feasibility of instructions generated by models, regardless of whether the instruction led the Builder towards task success. For instruction-type utterances, evaluators were asked to select one of the following categories:

- *Perfectly clear:* given the current state of the board, the instruction is clear enough such that it can be immediately executed by **B**; i.e., all references to blocks, shapes, colors, spatial relations, etc. in the utterance constitute a description that is consistent with and executable in the current game.

- *Somewhat unclear:* the blocks/features described in the instruction are consistent with the current game state, but the instruction itself is ambiguous or underspecified and is therefore not immediately executable.

- *Completely unclear or impossible:* the instruction describes blocks/features that are not consistent with the current game state, or is impossible to execute in the current game state.

[1] This category was originally split into two separate but very similar categories, *"Correct **B**'s actions or plans"* and *"Clarify or correct **A**'s descriptions or instructions"*. These categories were merged post-hoc after discovering that the ambiguity of the two definitions led to poor inter-annotator agreement on the individual categories.

Executability results are shown in Table 3 ($\alpha = 0.860$). While humans here have a lower rate of generating instructions (see Table 3 of the main paper), the instructions they do produce are almost always perfectly executable.

4.5 Correctness

In addition to the colors, spatial relations, and other entity properties mentioned in the utterance, evaluators were asked to rate the correctness of the utterance with respect to the target structure and overall task goal. Here, it is important to note that an utterance can be fully correct without necessarily needing to be immediately executable: e.g., *"We're going to build a row of 3 green blocks"* may not be specific enough for **B** to immediately take action (*maybe executable*), but it can be correct with respect to the target structure if said structure contained such a row that could feasibly placed at that point in the game (*fully correct*). Additionally, placement of temporary blocks may be necessary to eventually build "floating" (suspended) blocks; these could also be deemed correct invariant of color as long as their relative placement followed a reasonably efficient path towards the target.

Evaluators were asked to rate an utterance's correctness according to the following:

- *Fully correct:* all elements of the instruction that are described (colors, spatial relations) that should be a part of the final structure are consistent with the target.
- *Partially correct:* some elements of the instruction that are described (minimally, the type of action and the color of the block to be used) are correct with respect to the overall target structure, while other elements are incorrect. With some minor corrections to the utterance, the instruction can be seen as close to being fully correct with respect to the target, but slightly misses the mark.
- *Completely incorrect:* the elements described in the instruction are completely incorrect with respect to the target structure.
- *N/A:*[2] the utterance does not contain enough information to be judged for correctness.

Correctness results are shown in Table 4 of the main paper ($\alpha = 0.795$). Associated analysis and discussion can be found in Section 9.2.

5 Qualitative Analysis of Generated Utterances

Here, we provide a couple of examples of utterances generated by our model, placed within the context of the game state with accompanying screenshots.

Example 1. This example shows a game in which the target in Figure 3a is being built.

In this instance, the model generates *"and then put a yellow block on top of that"*. While the generated spatial relation is incorrect with respect to the target structure, the color of the mentioned block (yellow) is correct.

Example 2. This example shows a game in which the target in Figure 4a is being built. At this point in the game (Figure 4b), the Builder has not placed any blocks yet.

In this instance, the model generates *"place a red block on the ground"*. In this case, both the color of the mentioned block (red) and its spatial relation with respect to the ground are correct.

References

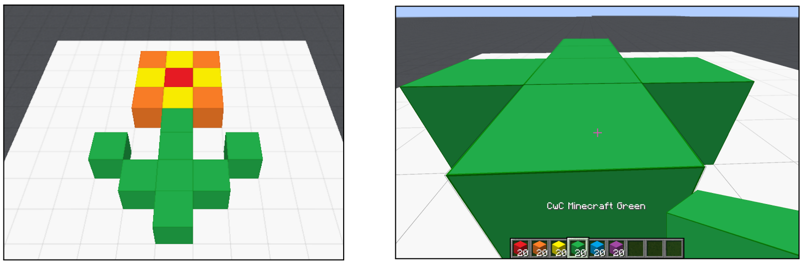
Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

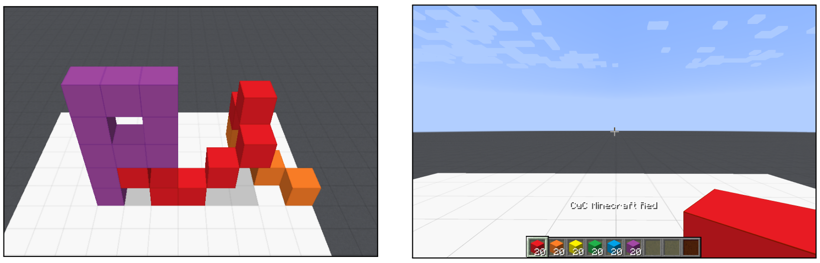
[2] Originally, *N/A* was used to indicate either that an utterance was not informative enough or that it had already been disqualified due to being an unclear/impossible instruction (see Section 4.4). Those utterances deemed non-executable and also marked as *N/A* were modified to be labeled as *Incorrect* in a postprocessing step.



(a) Target structure

(b) Game state

Figure 3: Example 1: target structure vs. current world state. The utterance generated here is *"and then put a yellow block on top of that"*, which in this context is partially correct (yellow blocks do need to be placed next, but not in the location described).



(a) Target structure

(b) Game state

Figure 4: Example 2: target structure vs. current world state. The utterance generated here is *"place a red block on the ground"*, which is fully correct.

A Term-Specific Word & Synonym Lists

We use the following word lists to compute term-specific metrics:

- **Colors:** red, orange, yellow, purple, green, blue
- **Spatial relations:** top, right, left, side, middle, up, down, bottom, towards, center, above, diagonal, out, front, here, away, diagonally, behind, back, between, below, vertical, long, tall, sides, flat, touching, high, facing, under, directly, opposite, toward, parallel, standing, near, forward, wide, horizontal, face, underneath, closest, across, perpendicular, rightmost, closer, along, leftmost, vertically, looking, around, whole, centered, degrees, extending, 90, 2d, before, sticking, topmost, edges, adjacent, mirror, perspective, attached, upside, highest, height, touch, upwards, hanging, straight, higher, big, shifted, inside, lower, horizontally, connecting, reference, orientation, upper, upright, inner, stacked, length, longer, apart, small, symmetric, furthest, float, upward, ahead, farthest, hole, hang, outward, angle, faces, short, 180, shorter, oriented, entire, outer, outside, outwards, overhanging, taller, symmetrical, jutting, beneath, inward, inwards, 3d, diagonals
- **Dialogue:** ?, ok, place, put, okay, make, good, sorry, yes, build, another, other, same, add, perfect, yeah, great, no, next, first, remove, last, done, yep, not, cool, nice, placed, stack, move, delete, yup, hello, hi, again, alright, connect, starting, ready, making, break, please, bad, extend, fill, yea, use, check, rid, ya, sure, awesome, correct, gotcha, repeat, leave, exactly, connected, yay, switch, keep, nah, shift, hey, enough, fine, thanks, complete, stand, replace, almost, excellent, oops, rotate, wrong, nope, leaving, punch, continue, finish, sweet, whoops, additional, mistake, placing, removed, final, thank, copy, turn, create, once

The synonym map used for generating additional references using synonym replacement is as follows:

Word	Synonyms
two	2
three	3
four	4
five	5

six	6
seven	7
eight	8
nine	9
second	2nd
2nd	second
third	3rd
3rd	third
okay	ok
ok	okay
put	place, add
place	put, add
add	put, place
placed	put, added
make	build
build	make
made	built
built	made
building	making
yes	yep, ya, yeah, yup, yea
yep	yes, ya, yeah, yup, yea
ya	yep, yes, yeah, yup, yea
yeah	yes, yep, ya, yup, yea
yup	yes, yep, ya, yeah, yea
yea	yes, yep, ya, yeah, yup
ground	floor
perfect	good, great, awesome, nice, cool, alright
great	good, perfect, awesome, nice, cool, alright
nice	good, great, awesome, perfect, cool, alright
cool	good, great, awesome, perfect, nice, alright
awesome	good, perfect, great, nice, cool, alright
alright	good, perfect, great, nice, cool, awesome
but	however
however	but
start	begin
begin	start
starting	beginning
beginning	starting
remove	delete, break
delete	remove, break
break	remove, delete
floating	hovering
hovering	floating
towards	toward
toward	towards
looks	seems
brick	block
block	brick
bricks	blocks
blocks	bricks
ones	blocks, bricks
done	finished
move	shift
shift	move
spaces	squares
squares	blocks, bricks
think	believe
believe	think
gap	space
shape	structure
hello	hi
hi	hello
empty	blank
blank	empty
below	under, underneath
under	below, underneath
underneath	below, under
tower	stack, pillar
pillar	stack, tower
connect	join
join	connect
connected	joined
joined	connected
level	layer
u	you
you	u
tall	high
high	tall
lol	haha
haha	lol
wait	stop
stop	wait
staircase	stairway
stairway	staircase
want	need
need	want
job	work, stuff
please	pls
pls	please
perpendicular	orthogonal
orthogonal	perpendicular
correct	right
gotcha	understood
understood	gotcha
repeat	redo, mimic
redo	repeat, mimic
mimic	redo, repeat
sort	kind
kind	sort
table	desk
desk	table
problem	worries, prob
worries	problem, prob
prob	problem, worries