

# On the Effects of Data Scale on UI Control Agents

**Wei Li<sup>1</sup>, William Bishop<sup>1</sup>, Alice Li<sup>1</sup>, Chris Rawles<sup>1</sup>,  
Folawiyo Campbell-Ajala<sup>1</sup>, Divya Tyamagundlu<sup>2</sup>, and Oriana Riva<sup>1</sup>**

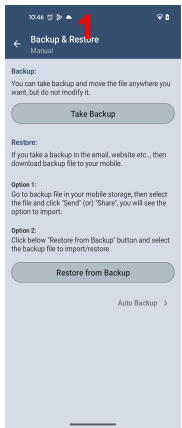
1. Google DeepMind

2. Google

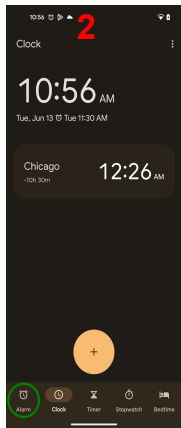
# The AndroidControl dataset

Dataset	Platform	# Human demos	# Unique instr.	# Apps or websites	# Task steps	UI tree?	Screen?	High-level instr.	Low-level instr.
MiniWoB++ [31]	Web (synthetic)	17,971	100	114	2.3	✓	✗	✗	✓
WebShop [40]	Web	1,566	1,566	1	11.3	✗	✓	✓	✗
UIBert [2]	Android	16,660	-	-	1.0	✓	✓	✗	✓
PixelHelp [22]	Android	187	187	4	4.2	✓	✗	✓	✓
UGIF [33]	Android	523	523	12	5.3	✓	✓	✓	✓
MoTIF [6]	Android	4,707	276	125	4.5	✓	✓	✓	✓
Mind2Web [7]	Web	2,350	2,350	137	7.3	✓	✓	✓	✗
AitW [27]	Android	715,142	30,378	357	6.5	✗	✓	✓	✗
WebVoyager [12]	Web	643	643	15	-	✓	✓	✓	✗
WebLINX [24]	Web	2,337	2,377	155	43.0	✓	✓	✓	✗
<b>ANDROIDCONTROL</b>	Android	15,283	<b>14,548</b>	<b>833</b>	5.5	✓	✓	✓	✓

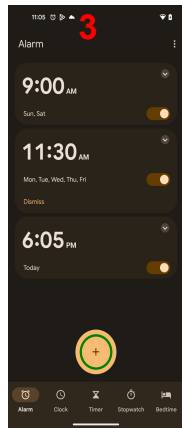
“In the clock app set an alarm for every Saturday at 6 am and called it time to walk”



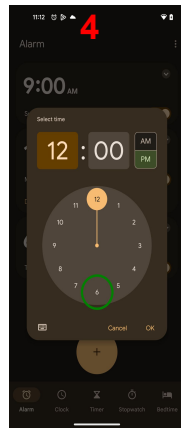
“Open Clock app”  
open\_app <deskclock>



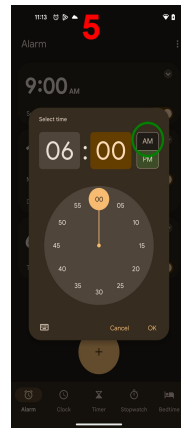
“Go to the alarm section”  
click <108,223>



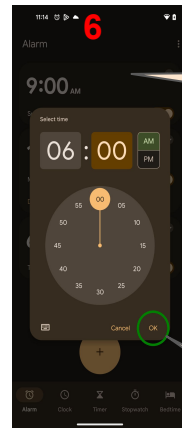
“Click on the add button”  
click <540,1959>



“Set hour to 6”  
click <541,1621>



“Click on the am”  
click <840,759>



“Click on OK option”  
click <866,1825>

high-level instruction

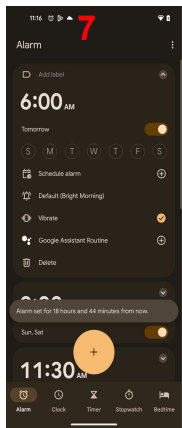
screenshot + accessibility tree

```
Package_name: "com.google.android.deskclock"  
View_id_resource_name: "com.google.android..."  
bounds_in_screen {  
  left: 782  
  top: 1762  
  right: 950  
  bottom: 1888  
}  
class_name: "android.widget.Button"  
text: "OK"  
content_description: ""  
hint_text: ""  
tooltip_text: ""  
is_checkable: false  
is_checked: false  
is_clickable: true
```

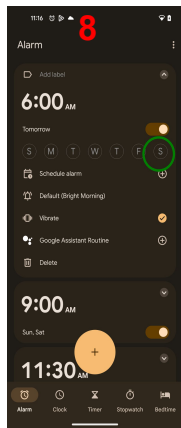
UI element metadata

low-level instruction

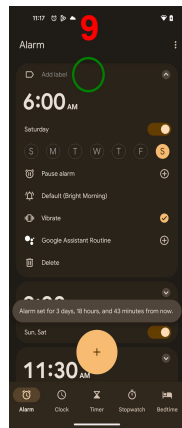
UI action



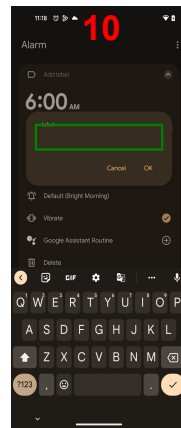
“Click on OK option”  
wait



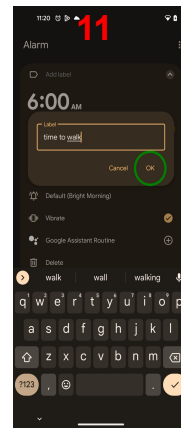
“Click on Saturday”  
click <855,820>



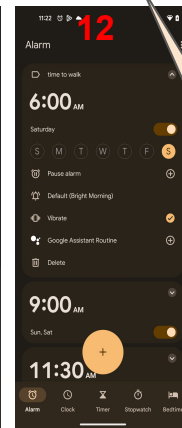
“Go to the label section”  
click <488,388>



“Name it time to walk”  
input\_text <"time to walk">

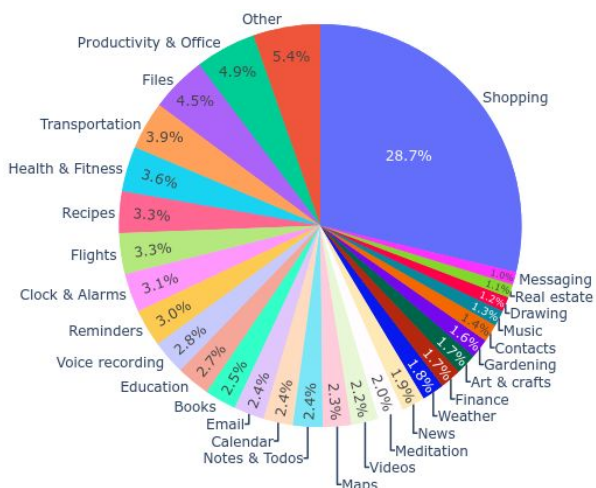


“Click the OK button”  
click <842,918>

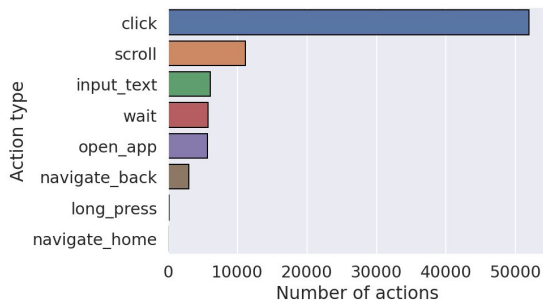


# Statistics

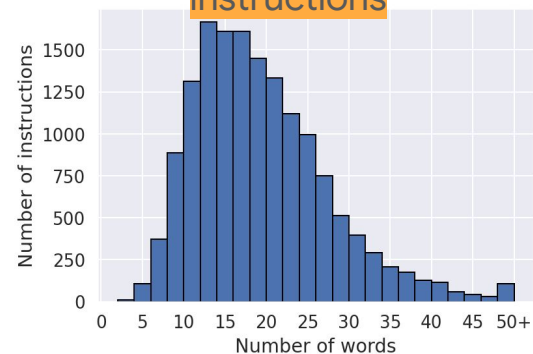
40 app categories



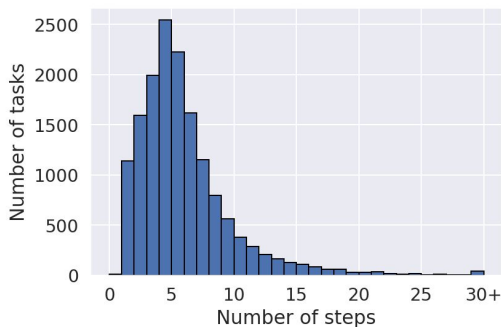
diverse actions



8-34 words in high-level instructions



tasks with 1-13 steps

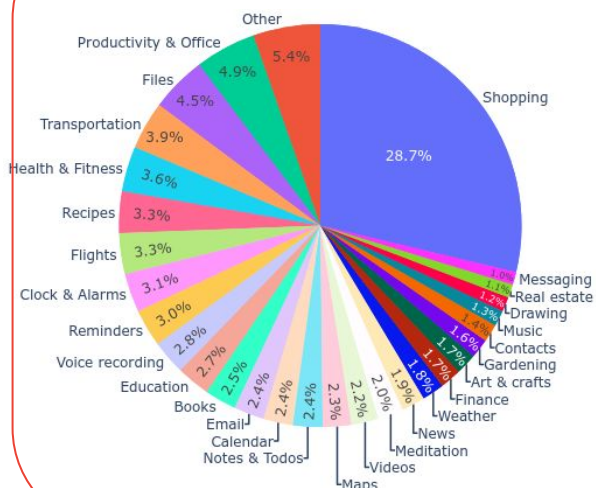


3-14 words in low-level instructions

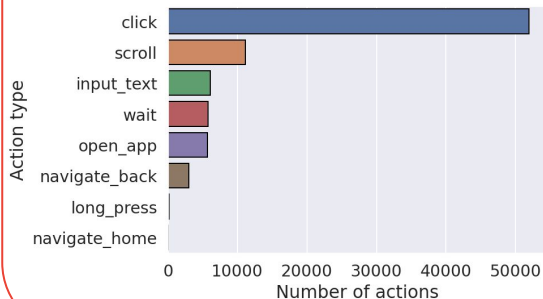


# Statistics

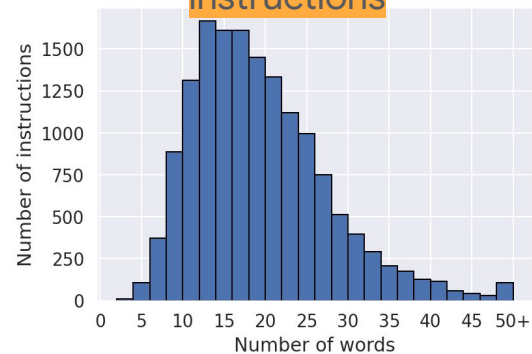
40 app categories



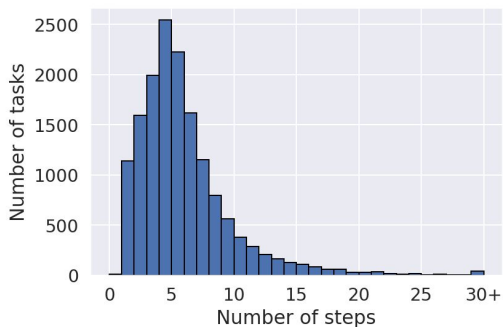
diverse actions



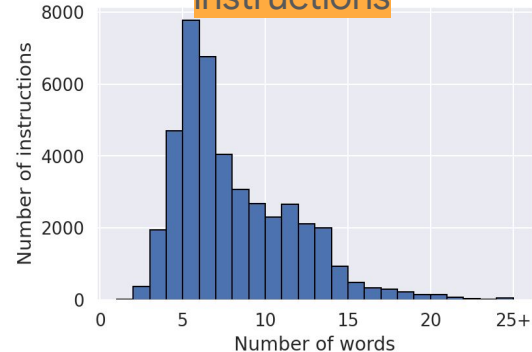
8-34 words in high-level instructions



tasks with 1-13 steps

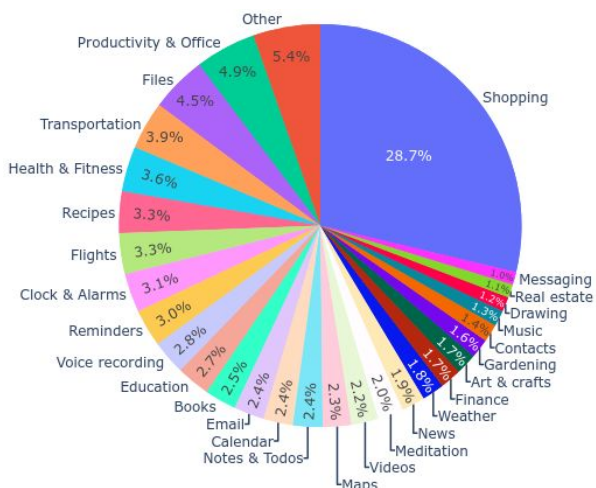


3-14 words in low-level instructions

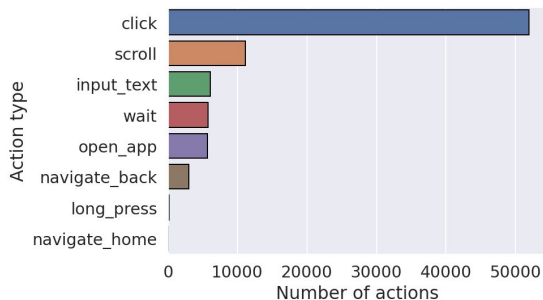


# Statistics

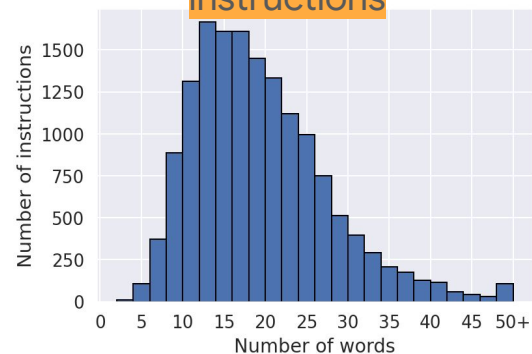
40 app categories



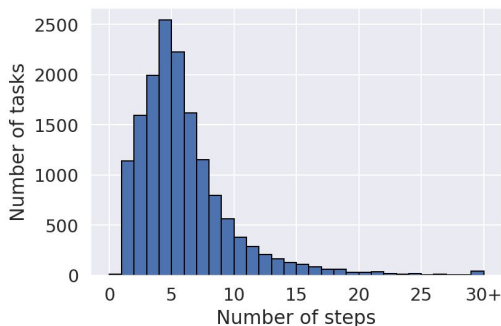
diverse actions



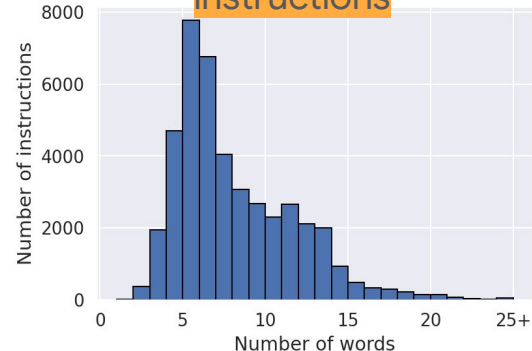
8-34 words in high-level instructions



tasks with 1-13 steps

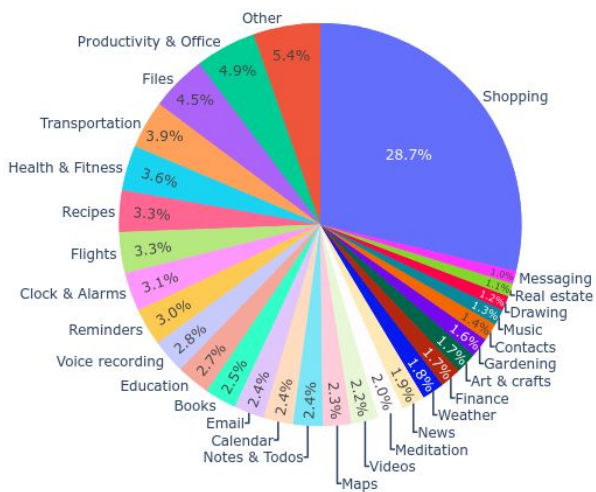


3-14 words in low-level instructions

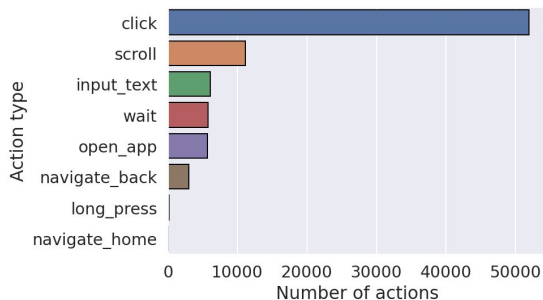


# Statistics

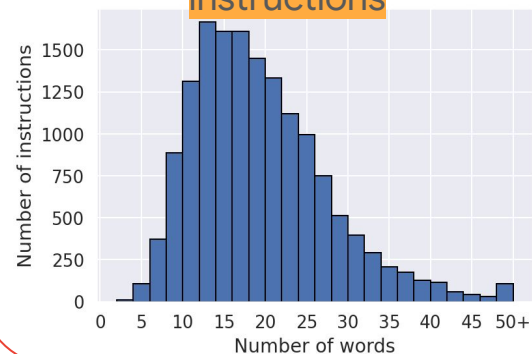
40 app categories



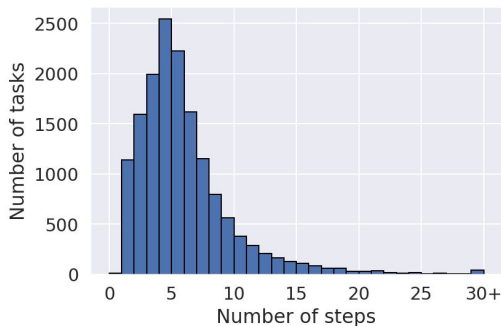
diverse actions



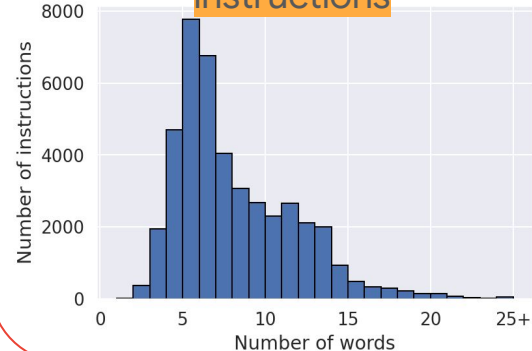
8-34 words in high-level instructions



tasks with 1-13 steps



3-14 words in low-level instructions



# AndroidControl splits

Split	Sub-splits	# Episodes	# Step actions	# Apps	# Categories	Avg. # elements per screen
Train	-	13,604	74,722	769	39	222.2
Val	-	137	690	99	29	214.4
Test	IDD	721	3,897	296	35	221.5
	App-unseen	631	3,475	64	12	185.4
	Task-unseen	803	4,464	90	12	181.6
	Category-unseen	700	3,891	68	4	184.7

in-domain test split

out-of-domain test split



## Performance on the in-domain sub-split

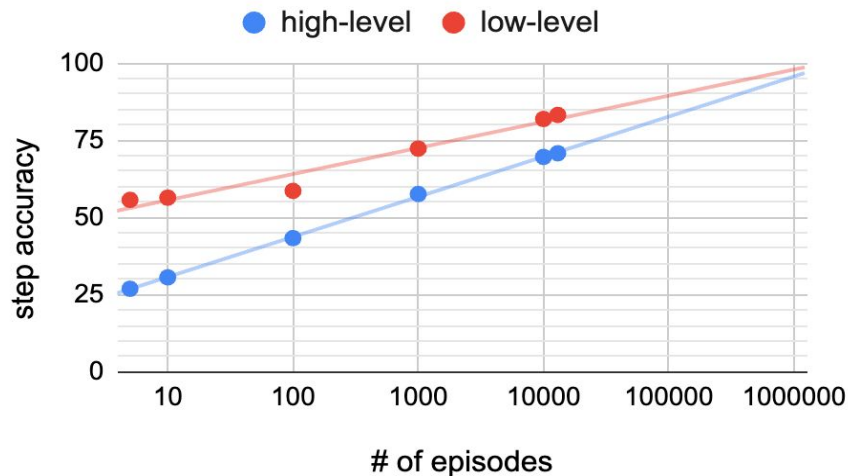
Regime	Method	Model	Step accuracy	
			high-level instr.	low-level instr.
Zero-shot	AitW	PaLM 2L	19.5	<b>56.7</b>
	SeeAct	GPT-4-Turbo	33.9	54.3
	M3A	GPT-4-Turbo	<b>42.1</b>	55.0
	ER	PaLM 2S	19.5	45.5
	ER	PaLM 2L	33.0	45.9
	ER	GPT-4	32.1	51.7
	ER	Gemini 1.5 Pro	24.4	50.2
Few-shot	FS-5	Gemini 1.5 Pro	<b>41.8</b>	50.2
	FS-10	Gemini 1.5 Pro	40.2	50.8
	FS-100	Gemini 1.5 Pro	39.5	<b>53.3</b>
LoRA-tuned	LT-5	PaLM 2S	30.3	57.1
	LT-10	PaLM 2S	28.5	58.9
	LT-100	PaLM 2S	39.8	62.8
	LT-1k	PaLM 2S	52.5	71.4
	LT-10k	PaLM 2S	62.0	85.7
	LT-all	PaLM 2S	65.6	81.8
	LT-1k-r64	PaLM 2S	54.8	76.6
	LT-10k-r64	PaLM 2S	69.6	81.9
	LT-all-r64	PaLM 2S	<b>71.5</b>	<b>86.6</b>

# Fine-tuning performance using AndroidControl (PaLM 2S LoRA-tuned)

TARGET: **99% step-wise accuracy** required to achieve 95% episode accuracy for a 5-step task

## In-domain

*target achieved with 2M episodes*

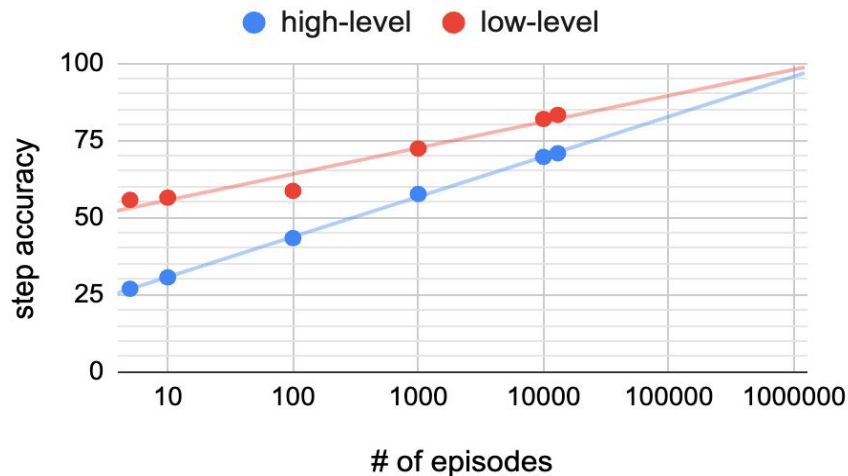


# Fine-tuning performance using AndroidControl (PaLM 2S LoRA-tuned)

TARGET: **99% step-wise accuracy** required to achieve 95% episode accuracy for a 5-step task

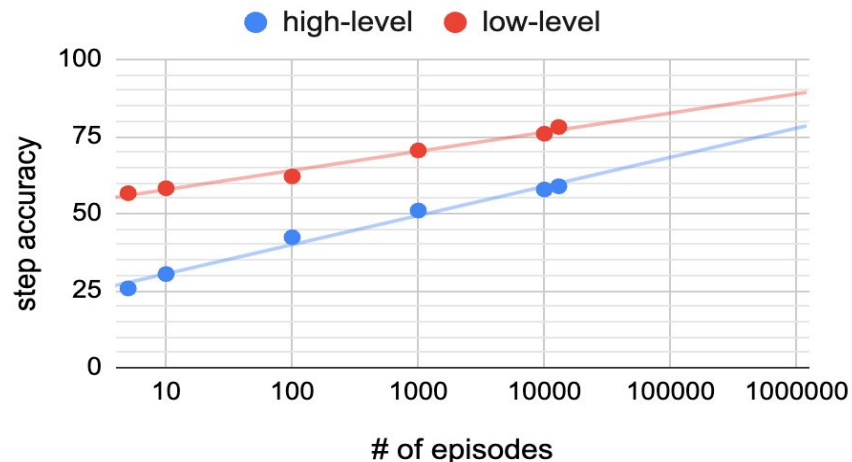
## In-domain

*target achieved with 2M episodes*



## Out-of-domain

*target achieved with 150M episodes*



# IDD vs. OOD

		IDD	app-unseen	task-unseen	categ-unseen
LT-5	HL	26.9	25.7 [-1.2]	26.4 [-0.5]	25.1 [-1.8]
	LL	55.7	56.9 [+1.2]	56.6 [+0.9]	56.4 [+0.7]
LT-10	HL	30.6	29.9 [-0.7]	31.1 [+0.5]	30.2 [-0.4]
	LL	56.4	58.3 [+1.9]	58.2 [+1.8]	58.2 [+1.8]
LT-100	HL	43.3	42.4 [-0.9]	42.5 [-0.8]	42.1 [-1.2]
	LL	58.6	62.7 [+4.1]	61.7 [+3.1]	61.8 [+3.2]
LT-1k	HL	53.2	49.0 [-4.2]	49.3 [-3.9]	48.1 [-5.1]
	LL	68.0	68.0 [ 0.0]	67.3 [-0.7]	67.4 [-0.6]
LT-10k	HL	63.9	55.2 [-8.7]	55.6 [-8.3]	54.2 [-7.7]
	LL	78.7	76.7 [-2.0]	75.6 [-3.1]	75.5 [-3.2]
LT-all	HL	65.5	58.7 [-6.8]	59.7 [-5.8]	58.2 [-7.3]
	LL	80.7	78.6 [-2.1]	77.9 [-2.8]	77.8 [-2.9]
LT-1k-r64	HL	57.6	51.1 [-6.5]	51.7 [-5.9]	50.2 [-7.4]
	LL	72.3	71.0 [-1.3]	70.4 [-1.9]	70.1 [-2.2]
LT-10k-r64	HL	69.6	57.7 [-11.9]	56.9 [-12.7]	58.9 [-10.7]
	LL	81.9	76.3 [-5.6]	75.8 [-6.1]	75.2 [-6.7]
LT-all-r64	HL	70.8	58.5 [-12.3]	59.6 [-11.2]	57.4 [-13.4]
	LL	83.2	78.5 [-4.7]	77.3 [-5.9]	76.8 [-6.4]

- IDD: in-domain split
- OOD: out-of-domain splits
  - app-unseen
  - task-unseen
  - Category-unseen
- HL: high-level instructions
- LL: low-level instructions
- LT-X: LoRA tuned on X training episodes with LoRA rank = 4
- LT-X-r64: LoRA tuned on X training episodes with LoRA rank = 64
- [-Y] percentage point decrease from IDD accuracy
- [+Y] percentage point increase from IDD accuracy

# IDD vs. OOD

		IDD	app-unseen	task-unseen	categ-unseen
LT-5	HL	26.9	25.7 [-1.2]	26.4 [-0.5]	25.1 [-1.8]
	LL	55.7	56.9 [+1.2]	56.6 [+0.9]	56.4 [+0.7]
LT-10	HL	30.6	29.9 [-0.7]	31.1 [+0.5]	30.2 [-0.4]
	LL	56.4	58.3 [+1.9]	58.2 [+1.8]	58.2 [+1.8]
LT-100	HL	43.3	42.4 [-0.9]	42.5 [-0.8]	42.1 [-1.2]
	LL	58.6	62.7 [+4.1]	61.7 [+3.1]	61.8 [+3.2]
LT-1k	HL	53.2	49.0 [-4.2]	49.3 [-3.9]	48.1 [-5.1]
	LL	68.0	68.0 [0.0]	67.3 [-0.7]	67.4 [-0.6]
LT-10k	HL	63.9	55.2 [-8.7]	55.6 [-8.3]	54.2 [-7.7]
	LL	78.7	76.7 [-2.0]	75.6 [-3.1]	75.5 [-3.2]
LT-all	HL	65.5	58.7 [-6.8]	59.7 [-5.8]	58.2 [-7.3]
	LL	80.7	78.6 [-2.1]	77.9 [-2.8]	77.8 [-2.9]
LT-1k-r64	HL	57.6	51.1 [-6.5]	51.7 [-5.9]	50.2 [-7.4]
	LL	72.3	71.0 [-1.3]	70.4 [-1.9]	70.1 [-2.2]
LT-10k-r64	HL	69.6	57.7 [-11.9]	56.9 [-12.7]	58.9 [-10.7]
	LL	81.9	76.3 [-5.6]	75.8 [-6.1]	75.2 [-6.7]
LT-all-r64	HL	70.8	58.5 [-12.3]	59.6 [-11.2]	57.4 [-13.4]
	LL	83.2	78.5 [-4.7]	77.3 [-5.9]	76.8 [-6.4]

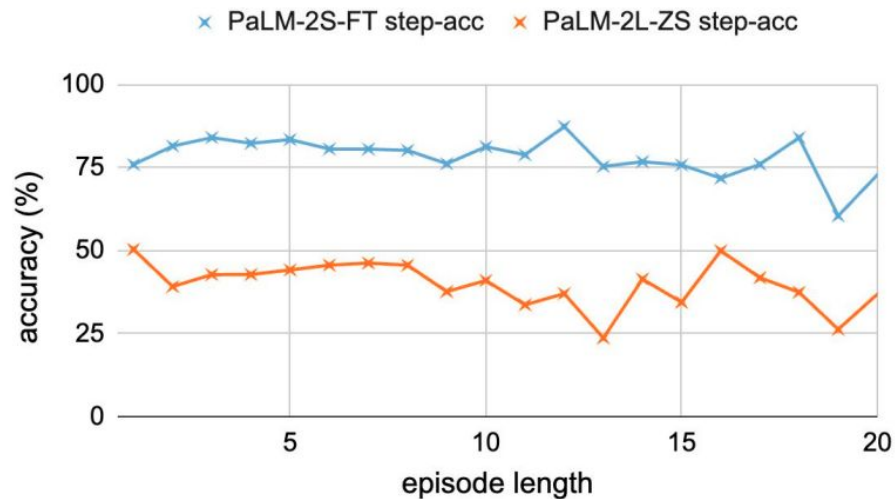
- IDD: in-domain split
- OOD: out-of-domain splits
  - app-unseen
  - task-unseen
  - Category-unseen
- HL: high-level instructions
- LL: low-level instructions
- LT-X: LoRA tuned on X training episodes with LoRA rank = 4
- LT-X-r64: LoRA tuned on X training episodes with LoRA rank = 64
- [-Y] percentage point decrease from IDD accuracy
- [+Y] percentage point increase from IDD accuracy

# IDD vs. OOD

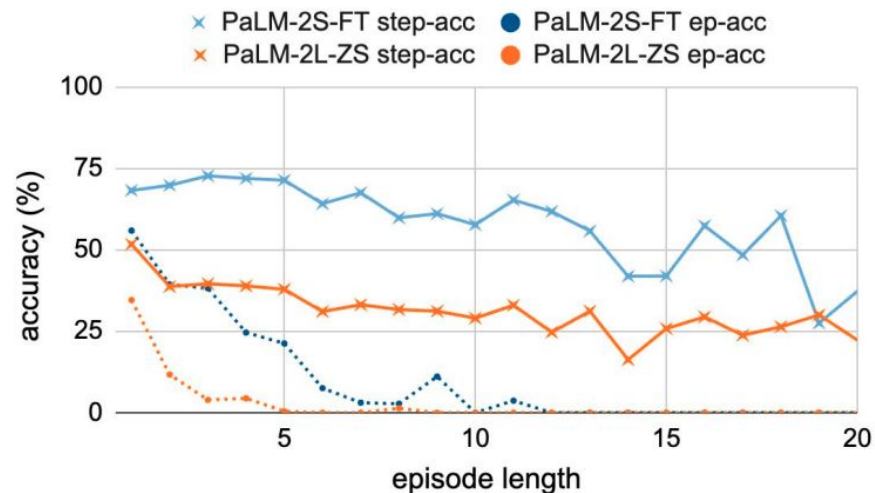
		IDD	app-unseen	task-unseen	categ-unseen
LT-5	HL	26.9	25.7 [-1.2]	26.4 [-0.5]	25.1 [-1.8]
	LL	55.7	56.9 [+1.2]	56.6 [+0.9]	56.4 [+0.7]
LT-10	HL	30.6	29.9 [-0.7]	31.1 [+0.5]	30.2 [-0.4]
	LL	56.4	58.3 [+1.9]	58.2 [+1.8]	58.2 [+1.8]
LT-100	HL	43.3	42.4 [-0.9]	42.5 [-0.8]	42.1 [-1.2]
	LL	58.6	62.7 [+4.1]	61.7 [+3.1]	61.8 [+3.2]
LT-1k	HL	53.2	49.0 [-4.2]	49.3 [-3.9]	48.1 [-5.1]
	LL	68.0	68.0 [ 0.0]	67.3 [-0.7]	67.4 [-0.6]
LT-10k	HL	63.9	55.2 [-8.7]	55.6 [-8.3]	54.2 [-7.7]
	LL	78.7	76.7 [-2.0]	75.6 [-3.1]	75.5 [-3.2]
LT-all	HL	65.5	58.7 [-6.8]	59.7 [-5.8]	58.2 [-7.3]
	LL	80.7	78.6 [-2.1]	77.9 [-2.8]	77.8 [-2.9]
LT-1k-r64	HL	57.6	51.1 [-6.5]	51.7 [-5.9]	50.2 [-7.4]
	LL	72.3	71.0 [-1.3]	70.4 [-1.9]	70.1 [-2.2]
LT-10k-r64	HL	69.6	57.7 [-11.9]	56.9 [-12.7]	58.9 [-10.7]
	LL	81.9	76.3 [-5.6]	75.8 [-6.1]	75.2 [-6.7]
LT-all-r64	HL	70.8	58.5 [-12.3]	59.6 [-11.2]	57.4 [-13.4]
	LL	83.2	78.5 [-4.7]	77.3 [-5.9]	76.8 [-6.4]

- IDD: in-domain split
- OOD: out-of-domain splits
  - app-unseen
  - task-unseen
  - Category-unseen
- HL: high-level instructions
- LL: low-level instructions
- LT-X: LoRA tuned on X training episodes with LoRA rank = 4
- LT-X-r64: LoRA tuned on X training episodes with LoRA rank = 64
- [-Y] percentage point decrease from IDD accuracy
- [+Y] percentage point increase from IDD accuracy

# Accuracy vs. episode length



(a) Low-level instructions



(b) High-level instructions

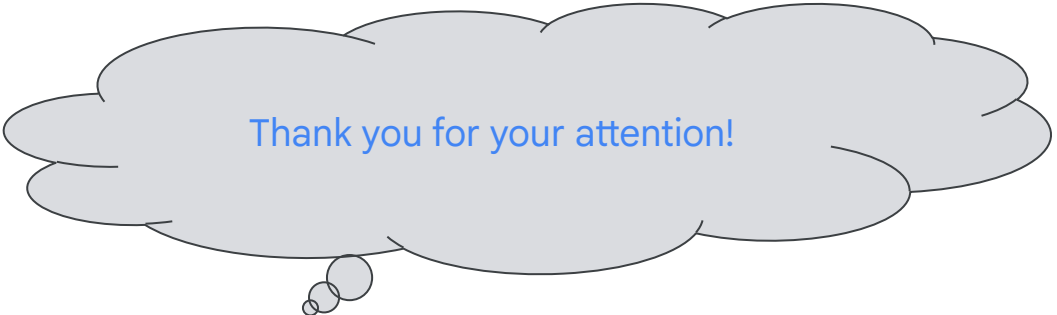
# Summary

- AndroidControl: a large and diverse dataset for studying the performance of UI agents in and out of domain
- The performance of fine-tuned models improves linearly to the number of training episodes
- A sufficiently fine-tuned model outperforms much larger models operating in zero-shot or few-shot setup
- Out of domain fine-tuning alone is not a viable solution because of the very large number of training data required to handle high-level instructions



# Summary

- AndroidControl: a large and diverse dataset for studying the performance of UI agents in and out of domain
- The performance of fine-tuned models improves linearly to the number of training episodes
- A sufficiently fine-tuned model outperforms much larger models operating in zero-shot or few-shot setup
- Out of domain fine-tuning alone is not a viable solution because of the very large number of training data required to handle high-level instructions



Thank you for your attention!