Appendix A. Dataset Description

The instruction induction dataset was proposed by Honovich et al. (2023) and contains 24 instruction induction tasks. As shown in Table A.1, eight tasks (First Letter, List Letters, Pluralization, Passivization, Larger Animal, Sum, Difference, Number to Word) were excluded because baseline methods have achieved near-perfect accuracy. Additionally, three tasks (Cause Selection, Common Concept, Formality) were excluded due to having fewer than 50 samples.

Table A.1 24 instruction induction tasks.

Category	Task	Description of task	Demonstration
Spelling	First Letter	Extract the first letter of the input word.	cat → c
	Second Letter	Extract the second letter of the input word.	cat → a
	List Letters	Break the input word into letters, separated by spaces.	cat → c a t
	Starting With	Extract the words starting with a given letter from the input sentence.	The man whose car I hit last week sued me. [m] → man, me
Morphosyntax	Pluralization	Convert the input word to its plural form.	cat → cats
	Passivization	Write the input sentence in passive form.	The artist introduced the scientist. → The scientist was introduced by the artist.
Syntax	Negation	Negate the input sentence.	Time is finite → Time is not finite

Table A.1 24 instruction induction tasks. (continued)

Category	Task	Description of task	Demonstration
Lexical Semantics	Antonyms	Write a word that means the opposite of the input word.	won → lost
	Synonyms	Write a word with a similar meaning to the input word.	alleged → supposed
	Membership	Write all the animals that appear in the given list.	cat, helicopter, cook, whale, frog, lion → frog, cat, lion, whale
Phonetics	Rhymes	Write a word that rhymes with the input word.	sing → ring
Knowledge	Larger Animal	Write the larger of the two given animals	koala, snail → koala
Semantics	Cause Selection	Find which of the two given cause and effect sentences is the cause.	Sentence 1: The soda went flat. Sentence 2: The bottle was left open. → The bottle was left open.
	Common Concept	Find a common characteristic for the given objects.	guitars, pendulums, neutrinos → involve oscillations.

Table A.1 24 instruction induction tasks. (continued)

Category	Task	Description of task	Demonstration
Style	Formality	Rephrase the sentence in formal language.	Please call once you get there → Please call upon your arrival.
Numerical	Sum	Sum the two given numbers.	22 10 → 32
	Difference	Subtract the second number from the first.	32 22 → 10
	Number to Word	Write the number in English words.	26 → twenty-six
Multilingual	Translation en-de	Translate the word into German.	game → spiel
	Translation en-es	Translate the word into Spanish.	game → juego
	Translation en-fr	Translate the word into French.	game → jeu
GLUE	Sentiment Analysis	Determine whether a movie review is positive or negative.	The film is small in scope, yet perfectly formed. → positive
	Sentence Similarity	Rate the semantic similarity of two input sentences on a scale of 0 - definitely not to 5 - perfectly.	Sentence 1: A man is smoking. Sentence 2: A man is skating. → 0 - definitely not

Table A.1 24 instruction induction tasks. (continued)

Category	Task	Description of task	Demonstration
	Word in Context	Determine whether an input word has the same meaning in the two input sentences.	Sentence 1: Approach a task. Sentence 2: To approach the city. Word: approach → not the same

The counter factual tasks are derived from Wu et al. (2024). Following the study by Ye et al. (2024), we selected three types of tasks for our experiments: arithmetic, chess, and grammar, as detailed in Table A.2.

Table A.2 Counter factual evaluation tasks.

Task	Description of task	Demonstration
Arithmetic: Base-8	Add the two numbers in base-8.	76+76 → 174
Arithmetic: Base-9	Add the two numbers in base-9.	76+14 → 101
Arithmetic: Base-11	Add the two numbers in base-11.	76+14 → 8A
Arithmetic: Base-16	Add the two numbers in base-16.	EC+DD → 1C9

Table A.2 Counter factual evaluation tasks. (continued)

Task	Description of task	Demonstration
Chess: Swapping bishops and knights	Swap the initial positions of the knight and bishop in chess, then check if the first four moves in the input are legal. If each move is legal, output ”legal”; otherwise, output ”illegal”.	1. g3 Ng6 2. b3 Kf8 * → legal
Syntax: SOV	The structure of the input sentence is subject-object-verb. Find the main verb and the main subject in the sentence.	he good control has . → he has
Syntax: VSO	The structure of the input sentence is verb-subject-object. Find the main verb and the main subject in the sentence.	has he good control . → he has
Syntax: VOS	The structure of the input sentence is verb-object-subject. Find the main verb and the main subject in the sentence.	has good control he . → he has

Table A.2 Counter factual evaluation tasks. (continued)

Task	Description of task	Demonstration
Syntax: OVS	The structure of the input sentence is object-verb-subject. Find the main verb and the main subject in the sentence.	good control has he . → he has
Syntax: OSV	The structure of the input sentence is object-subject-verb. Find the main verb and the main subject in the sentence.	good control he has . → he has

Appendix B. Meta Prompts

1. Initialize instruction chain

You have a task to { instruction }.
List the steps to perform the task in order.
Please format the steps as follows inJSON:
{ output_format }

2. Forward propagation

You have a task to { instruction }.
Input:
{ input_text }
Perform the task on the input text by following steps:
{ instruction chain }
Please format as follows in JSON:
{ output_format }

3. Analyze the causes of errors

You have a task to { instruction }.
Input:
{ input_text }

You perform this task on the input in the following sequence of steps:
{ instruction_chain }
The output is:
{ model_output }
But the real answer should be:
{ answer }
Analyze the reasons for incorrect output when the steps are followed.
Please format the reason (or reasons) as follows in JSON:
{ output_format }

4. Formulate improvement strategies
You have a task to { instruction }.
Input:
{ input_text }
You perform this task on the input in the following sequence of steps:
{ instruction_chain }
The output is:
{ model_output }
But the real answer should be:
{ answer }
The reason for following the steps but the output error are:
{ reasons }
Based on the above information, for each reason, analyze which step caused the error. And finally propose a strategy (or strategies) to improve the step.
Please format the strategy (or strategies) as follows in JSON:
{ output_format }

5. Update instruction chain
You have a task to { instruction }.
You list the steps to perform the task in order:
{ steps }
But there are drawbacks to the steps. Combine the following improvement strategies into the steps:
{ suggestions }
Please format the refined steps as follows in JSON:
{ output format }

References

Honovich, O., Shaham, U., Bowman, S.R., Levy, O., 2023. Instruction induction: From few examples to natural language task descriptions, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 1935–1952.

Wu, Z., Qiu, L., Ross, A., Akyrek, E., Chen, B., Wang, B., Kim, N., Andreas, J., Kim, Y., 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1819–1862.

Ye, Q., Axmed, M., Pryzant, R., Khani, F., 2024. Prompt engineering a prompt engineer. arXiv preprint. `arXiv:2311.05661`.