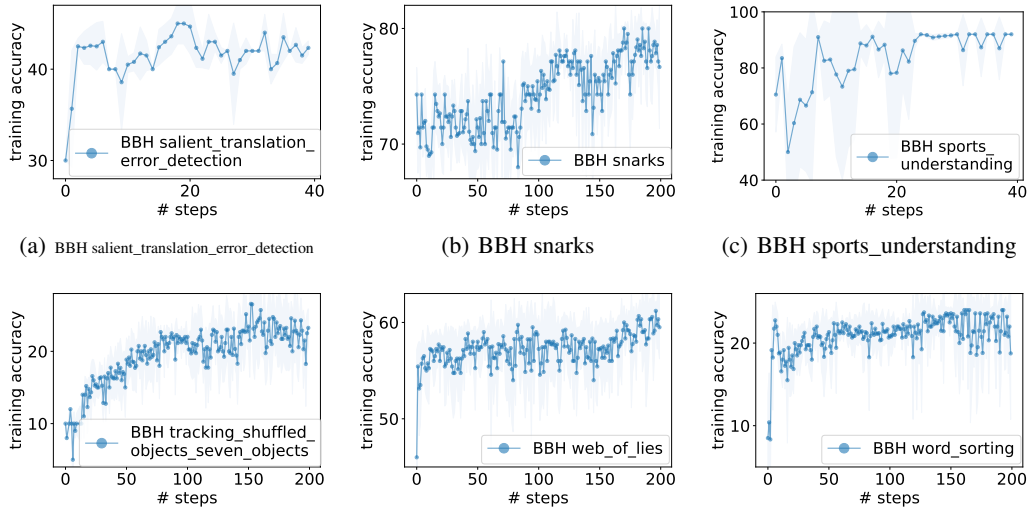

(a) BBH salient_translation_error_detection (b) BBH snarks (c) BBH sports_understanding

(d) BBH tracking_shuffled_objects_seven_objects (e) BBH web_of_lies (f) BBH word_sorting

Figure 24: Prompt optimization on 21 BBH tasks (except ruin_names and temporal_sequences in Figure 6) with the `text-bison` scorer and the `PaLM 2-L-IT` optimizer, Part II. All curves have upward trends.

E Prompt Optimization on BBH Tasks – Tabulated Accuracies and Found Instructions

E.1 `PaLM 2-L-IT` as optimizer, optimization starting from the empty string

Table 8 and 9 show the instructions found by prompt optimization. A comparison of their accuracies with baselines "Let's think step by step." (Kojima et al., 2022), "Let's work this out in a step by step way to be sure we have the right answer." (Zhou et al., 2022b), and the empty string is in Table 7; a visualization is in Section 5.2 Figure 5.

Table 7: Accuracies on BBH tasks: our found instructions with the PaLM 2-L-IT optimizer vs baseline. The optimization starts from the empty string. Because of the 20-80 train-test split, we show accuracies with the format "training / test / overall (training + test)". The PaLM 2-L scores are from A_begin instructions; the text-bison scores are from Q_begin instructions. Bold numbers indicate the best for the corresponding task.

Task	Scorer	Our Acc	"Let's think step by step." Acc	"Let's work this out in a step by step way to be sure we have the right answer." Acc	empty string "" Acc
		training / test / overall	training / test / overall	training / test / overall	training / test / overall
boolean_expressions	PaLM 2-L	**90.0 / 83.5 / 84.8**	90.0 / 83.0 / 84.4	82.0 / 74.0 / 75.6	74.0 / 71.0 / 71.6
causal_judgement	PaLM 2-L	**84.8 / 58.0 / 63.1**	73.0 / 55.3 / 58.8	59.5 / 57.3 / 57.8	29.7 / 49.3 / 45.5
date_understanding	PaLM 2-L	**86.0 / 84.5 / 84.8**	76.0 / 80.0 / 79.2	74.0 / 77.0 / 76.4	70.0 / 74.0 / 73.2
disambiguation_qa	PaLM 2-L	**80.0 / 69.0 / 71.2**	40.0 / 52.5 / 50.0	48.0 / 47.0 / 47.2	54.0 / 57.5 / 56.8
dyck_languages	PaLM 2-L	**100.0 / 100.0 / 100.0**	96.0 / 94.5 / 94.8	100.0 / 93.5 / 94.8	94.0 / 95.0 / 94.8
formal_fallacies	PaLM 2-L	**84.0 / 64.0 / 68.4**	78.0 / 59.5 / 63.2	68.0 / 63.0 / 64.0	66.0 / 59.0 / 60.4
geometric_shapes	PaLM 2-L	**76.0 / 57.0 / 60.8**	42.0 / 33.0 / 34.8	42.0 / 32.0 / 34.0	34.0 / 33.0 / 33.2
hyperbaton	PaLM 2-L	**100.0 / 96.0 / 96.8**	78.0 / 75.0 / 75.6	74.0 / 72.5 / 72.8	88.0 / 89.0 / 88.8
logical_deduction_seven_objects	PaLM 2-L	**74.0 / 57.0 / 60.4**	46.0 / 37.0 / 38.8	34.0 / 30.5 / 31.2	46.0 / 45.5 / 45.6
movie_recommendation	PaLM 2-L	**92.0 / 90.5 / 90.8**	62.0 / 52.5 / 54.4	52.0 / 48.0 / 48.8	80.0 / 83.0 / 82.4
multistep_arithmetic_two	PaLM 2-L	**72.0 / 55.5 / 58.8**	42.0 / 46.0 / 45.2	60.0 / 50.5 / 52.4	4.0 / 3.5 / 3.6
navigate	PaLM 2-L	**92.0 / 75.0 / 78.4**	68.0 / 62.0 / 63.2	70.0 / 64.0 / 65.2	38.0 / 37.5 / 37.6
object_counting	PaLM 2-L	**84.0 / 86.5 / 86.0**	36.0 / 46.5 / 44.4	60.0 / 62.0 / 61.6	28.0 / 27.0 / 27.2
penguins_in_a_table	PaLM 2-L	**86.2 / 71.8 / 74.7**	79.3 / 64.1 / 67.1	62.1 / 58.1 / 58.9	72.4 / 69.2 / 69.9
reasoning_about_colored_objects	PaLM 2-L	**98.0 / 85.5 / 88.0**	82.0 / 79.5 / 80.0	82.0 / 75.0 / 76.4	42.0 / 35.0 / 36.4
ruin_names	PaLM 2-L	**88.0 / 88.0 / 88.0**	70.0 / 55.0 / 58.0	80.0 / 75.5 / 76.4	88.0 / 76.5 / 78.8
salient_translation_error_detection	PaLM 2-L	**62.0 / 67.0 / 66.0**	42.0 / 50.0 / 48.4	58.0 / 46.0 / 48.4	56.0 / 56.5 / 56.4
snarks	PaLM 2-L	**85.7 / 83.2 / 83.7**	60.0 / 62.2 / 61.8	54.3 / 53.1 / 53.4	51.4 / 60.1 / 58.4
sports_understanding	PaLM 2-L	**98.0 / 88.0 / 90.0**	50.0 / 46.5 / 47.2	60.0 / 52.5 / 54.0	52.0 / 41.5 / 43.6
temporal_sequences	PaLM 2-L	**100.0 / 100.0 / 100.0**	100.0 / 96.0 / 96.8	90.0 / 87.0 / 87.6	100.0 / 99.5 / 99.6
tracking_shuffled_objects_seven_objects	PaLM 2-L	32.0 / 16.5 / 19.6	**58.0 / 61.5 / 60.8**	54.0 / 55.5 / 55.2	14.0 / 23.5 / 21.6
web_of_lies	PaLM 2-L	**62.0 / 52.0 / 54.0**	46.0 / 41.5 / 42.4	24.0 / 31.0 / 29.6	**54.0 / 54.0 / 54.0**
word_sorting	PaLM 2-L	**54.0 / 54.5 / 54.4**	2.0 / 4.5 / 4.0	12.0 / 9.5 / 10.0	20.0 / 22.5 / 22.0
boolean_expressions	text-bison	**98.0 / 87.0 / 89.2**	72.0 / 61.5 / 63.6	88.0 / 78.0 / 80.0	80.0 / 68.5 / 70.8
causal_judgement	text-bison	**78.4 / 58.0 / 62.0**	70.3 / 50.7 / 54.5	73.0 / 55.3 / 58.8	**78.4 / 58.0 / 62.0**
date_understanding	text-bison	**60.0 / 50.0 / 52.0**	44.0 / 45.5 / 45.2	48.0 / 45.0 / 45.6	44.0 / 45.0 / 44.8
disambiguation_qa	text-bison	**68.0 / 73.0 / 72.0**	4.0 / 6.0 / 5.6	4.0 / 15.5 / 13.2	52.0 / 68.5 / 65.2
dyck_languages	text-bison	**100.0 / 100.0 / 100.0**	100.0 / 95.5 / 96.4	100.0 / 94.5 / 95.6	100.0 / 98.5 / 98.8
formal_fallacies	text-bison	70.0 / 53.0 / 56.4	64.0 / 54.5 / 56.4	**84.0 / 82.5 / 82.8**	70.0 / 54.5 / 57.6
geometric_shapes	text-bison	**40.0 / 19.5 / 23.6**	22.0 / 13.0 / 14.8	18.0 / 12.0 / 13.2	20.0 / 14.5 / 15.6
hyperbaton	text-bison	**80.0 / 79.5 / 79.6**	64.0 / 67.5 / 66.8	64.0 / 69.0 / 68.0	64.0 / 64.0 / 64.0
logical_deduction_seven_objects	text-bison	66.0 / 53.5 / 56.0	**56.0 / 58.0 / 57.6**	56.0 / 56.0 / 56.0	58.0 / 56.5 / 56.8
movie_recommendation	text-bison	**98.0 / 90.0 / 91.6**	68.0 / 63.0 / 64.0	66.0 / 62.0 / 62.8	68.0 / 64.0 / 64.8
multistep_arithmetic_two	text-bison	**32.0 / 16.5 / 19.6**	12.0 / 18.0 / 16.8	18.0 / 17.5 / 17.6	16.0 / 18.5 / 18.0
navigate	text-bison	**72.0 / 61.0 / 63.2**	56.0 / 55.0 / 55.2	60.0 / 56.5 / 57.2	56.0 / 57.0 / 56.8
object_counting	text-bison	**72.0 / 62.0 / 64.0**	58.0 / 57.0 / 57.2	62.0 / 55.5 / 56.8	50.0 / 57.0 / 55.6
penguins_in_a_table	text-bison	**72.4 / 56.4 / 59.6**	58.6 / 53.0 / 54.1	55.2 / 55.6 / 55.5	58.6 / 53.0 / 54.1
reasoning_about_colored_objects	text-bison	**82.0 / 77.0 / 78.0**	76.0 / 72.5 / 73.2	78.0 / 73.0 / 74.0	74.0 / 69.5 / 70.4
ruin_names	text-bison	**88.0 / 82.5 / 83.6**	66.0 / 65.5 / 65.6	66.0 / 62.5 / 63.2	64.0 / 66.0 / 65.6
salient_translation _error_detection	text-bison	**46.0 / 50.5 / 49.6**	42.0 / 47.5 / 46.4	42.0 / 49.5 / 48.0	44.0 / 50.0 / 48.8
snarks	text-bison	**80.0 / 81.8 / 81.5**	68.6 / 77.6 / 75.8	71.4 / 76.2 / 75.3	77.1 / 84.6 / 73.1
sports_understanding	text-bison	**94.0 / 82.5 / 84.8**	86.0 / 79.0 / 80.4	90.0 / 81.0 / 82.8	38.0 / 44.5 / 43.2
temporal_sequences	text-bison	**78.0 / 81.0 / 80.4**	36.0 / 43.5 / 42.0	32.0 / 45.0 / 42.4	36.0 / 43.0 / 41.6
tracking_shuffled_objects_seven_objects	text-bison	**32.0 / 15.5 / 18.8**	10.0 / 17.0 / 15.6	10.0 / 18.0 / 16.4	12.0 / 15.5 / 14.8
web_of_lies	text-bison	**62.0 / 50.0 / 52.4**	48.0 / 45.5 / 46.0	48.0 / 44.0 / 44.8	52.0 / 51.5 / 51.2
word_sorting	text-bison	**24.0 / 17.5 / 18.8**	10.0 / 12.0 / 11.6	4.0 / 8.0 / 7.2	4.0 / 7.5 / 6.8

Table 8: BBH task-wise instructions found by prompt optimization with the PaLM 2-L scorer and the PaLM 2-L-IT optimizer. The optimization starts from the empty string.

Task	Our Instruction
boolean_expressions	A Boolean expression is a well-formed expression consisting of variables, values, and logical operators. The expression must evaluate to a single True or False value. The order of precedence of the logical operators is as follows: NOT, AND, OR, XOR, IMP. Parentheses can be used to group subexpressions and to control the order of evaluation.
causal_judgement	When considering questions about causation, a typical person would consider the following factors: whether the action or event was a necessary condition for the outcome to occur, a sufficient condition, a proximate cause, or a foreseeable cause.
date_understanding	To find the date X time ago from today, first find today's date. Then subtract X time from today's date. If the current date is the last day of a month, then the date a month ago is the last day of the previous month. If the current date is not the last day of a month, then the date a month ago is the same day of the previous month. For example, if today is March 31, 2023, then the date a month ago is February 28, 2023. If today is April 1, 2023, then the date a month ago is March 1, 2023.
disambiguation_qa	Identifying Antecedents of Pronouns: A Comprehensive Guide
dyck_languages	First, look for the opening parentheses. Then, count the number of opening parentheses. Finally, close the parentheses in the reverse order that they were opened.
formal_fallacies	A deductive argument is one where the conclusion follows necessarily from the premises. If the premises are true, then the conclusion must also be true. An invalid argument is one where it is possible for the premises to be true and the conclusion to be false.
geometric_shapes	A closed polygonal chain is a series of connected line segments. The line segments can be straight or curved. The first and last line segments are connected. The line segments do not intersect each other except at their endpoints. A closed polygon can be described by an SVG path element, which starts at a given point, goes to one or more additional points, and then ends at the starting point. The path element can consist of straight line segments, curved segments, or a mixture of both.
hyperbaton	The correct adjective order in English is opinion, size, shape, age, color, origin, material, and purpose. If you have more than one adjective of the same type, they are usually placed in order of importance. For example, you would say "a large, old, Pakistani ship" rather than "an old, large, Pakistani ship." There are a few exceptions to these rules, but they are generally followed in most cases.
logical_deduction _seven_objects	The following questions will test your ability to use deductive reasoning. You will be given a set of statements about a group of objects. You will then be asked to answer questions about the objects based on the statements. The statements in the questions are logically consistent, so you can use them to deduce the order of the objects. For each question, you must choose the option that is logically consistent with the information in the questions.
movie_recommendation	Based on your input, I have analyzed the given movies in terms of genre, plot, tone, audience rating, year of release, director, cast, and reviews. I have also taken into account the given options. The movie that is most similar to the given movies in terms of all these factors is:
multistep_arithmetic _two	The order of operations in mathematics is PEMDAS, which stands for Parentheses, Exponents, Multiplication, Division, Addition, and Subtraction. When there are multiple operations of the same precedence, they must be performed from left to right. Note that multiplication and division have the same precedence, as do addition and subtraction.
navigate	You will return to the starting point if and only if (1) the total number of steps you take forward is equal to the total number of steps you take back, and (2) the total number of turns you make is a multiple of 180 degrees.
object_counting	Here is a list of the objects you mentioned and their corresponding counts:
penguins_in_a_table	Here is my new text:
reasoning_about _colored_objects	Starting from the leftmost object in the row, I observe the following objects arranged in this order:
ruin_names	Which is the funniest pun on the artist or movie name?
salient_translation _error_detection	Instructions: Read the German sentence and its English translation carefully, then identify the type of error in the translation and select the correct option. There are six possible types of errors: Named Entities, Numerical Values, Modifiers or Adjectives, Negation or Antonyms, Facts, and Dropped Content.
snarks	Identify the sarcastic statement by considering the following factors: incongruity, exaggeration, understatement, context, speaker's intent, and audience's reaction. I will also consider the speaker's tone of voice, facial expressions, and body language.
sports_understanding	I will determine if a sentence about an athlete is plausible by first checking if it is grammatically correct. If it is, I will then check if it is consistent with the athlete's sport, position, and real-world statistics. I will also check if it is consistent with the rules of the athlete's sport. If the sentence is consistent with all of these things, I will answer "yes", otherwise I will answer "no".
temporal_sequences	The answer is the time that is not mentioned in the given statements.
tracking_shuffled_objects _seven_objects	Claire has the blue ball, Gertrude has the black ball, and Dave has the green ball. They are all happy with their new balls.
web_of_lies	The answer to a question is yes if there are an odd number of liars before the current speaker, and no if there are an even number of liars before the current speaker. If the current speaker is a truth-teller, they will say the opposite of what the previous person said, while a liar will say the same thing as the previous person said.
word_sorting	Alphabetical order of given words: