“Optimal” Prompt Engineering using Genetic Algorithms

Project Review for EN.705.651

Sridhar Sharma
Alejandro Salamanca
Son Pham

Agenda

- Quick Introduction
- Approach 1: Details and Results (Sridhar)
- Approach 2: Details and Results (Alejandro)
- Approach 3: Details and Results (Son)
- Development Challenges & Next Steps
- Previous work & References

Introduction

- Prompt Engineering is a new paradigm created due to the advent of Large Language models
- Entails developing and optimizing prompts for efficient usage of Large Language models (LLMs).
- Enhances LLMs' ability to handle complex tasks such as Question/Answering and Arithmetic reasoning
- Premise is that Prompting is critical for the performance of an LLM.
- Current prompting techniques are manual
- **Goals of the project**
  - Attempt to automate the generation of an “optimal” prompt using evolutionary (genetic algorithms).
  - Insight into optimal prompt design patterns
  - Create a metric to measure the effectiveness of a prompt.

Dataset

- GSM8K Dataset
  - 8.5K high quality linguistically diverse grade school math word problems created by human problem writers.
  - Train: 7.47k | Test: 1.32k
  - These problems take between 2 and 8 steps to solve, and solutions primarily involve performing a sequence of elementary calculations using basic arithmetic operations (+ - ×÷) to reach the final answer.

Input	**Output**
`Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?`	`Natalia sold 48/2 = <<48/2=24>>24 clips in May. Natalia sold 48+24 = <<48+24=72>>72 clips altogether in April and May. #### 72`

Baseline performance to beat

Num examples	Mistral 7B	Llama 13B	Llama 70B	GPT-3.5
0 (zero shot)	5%	4%	10%	23%
1 (one shot)	5%	13%	18%	36%
3 (few shot)	25%	21%	25%	54%

- Performance is measured on the base prompt: **Solve the math word problem, giving your answer as an arabic numeral**.

Genetic Algorithm

- Useful for Optimization problems with a “Fitness metric”
- Works on the survival of the “fittest” principle
- Randomly create the initial population
- Picking viable candidates that are carried forward to the next generation
- fitness score as well other techniques such as mutation and crossover generate members of the next population.
- Evolve the population over N generations

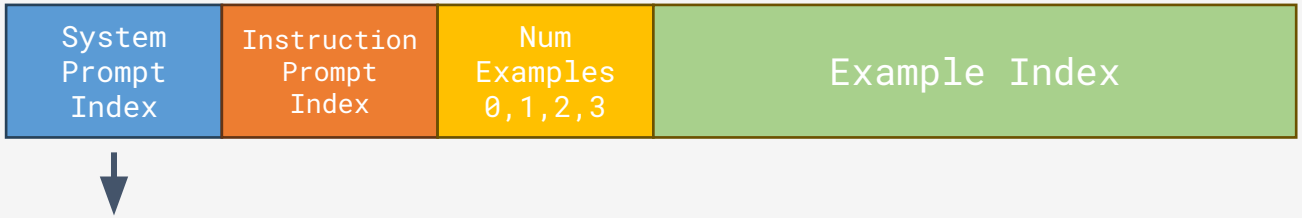
Approach 1

System Prompt & Instruction Prompt

Gene and Genetic Algorithm

- 19 System Prompts
- 14 instruction prompts
- Maximum Number of examples (3 and 8)- Random number
- 10 Members of population/generation
- 10 generations
- Best 60% of population moves to next generation
- 20% of population undergoes mutation
- 20% of the best members create children via crossover operation
- Each member evaluated with 10 problems from GSM8K Test data set
    - Same 10 problem set across 10 generations- overfitting
    - Different problems each generation
    - Cumulative Accuracy
    - Change fitness score to include number of problems
      fitness_score=cumulative_accuracy+ α * num_correct_responses

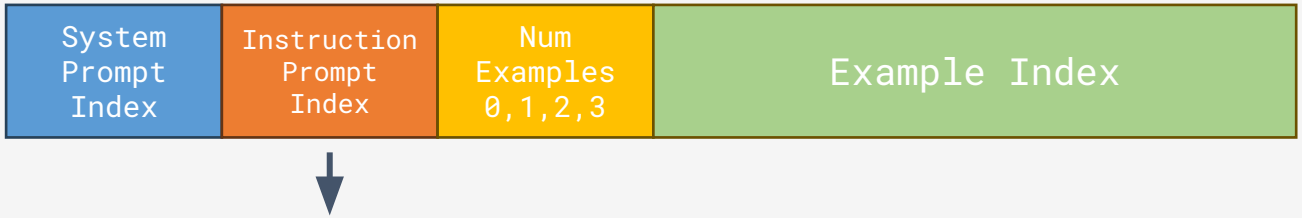
Gene and System Prompt




```
'Think like a graduate student and answer.',
'Think like a mathematics teacher and answer',
'Approach the problem as a mathematics educator and provide your
solution',
'Demonstrate the problem-solving process with the approach of a
math instructor',
'Offer your answer using the teaching methodology of a
mathematics educator',
'Think pedagogically and present your solution as if you were
teaching it in a classroom',
'Approach the solution with the clarity and explanation of a
mathematics teacher',
'Share your answer, providing step-by-step guidance akin to a
math educator',
'Explain your solution as if you were guiding students through
the problem as a math teacher'.
```

System prompts variants are derived using ChatGPT from some basic prompts

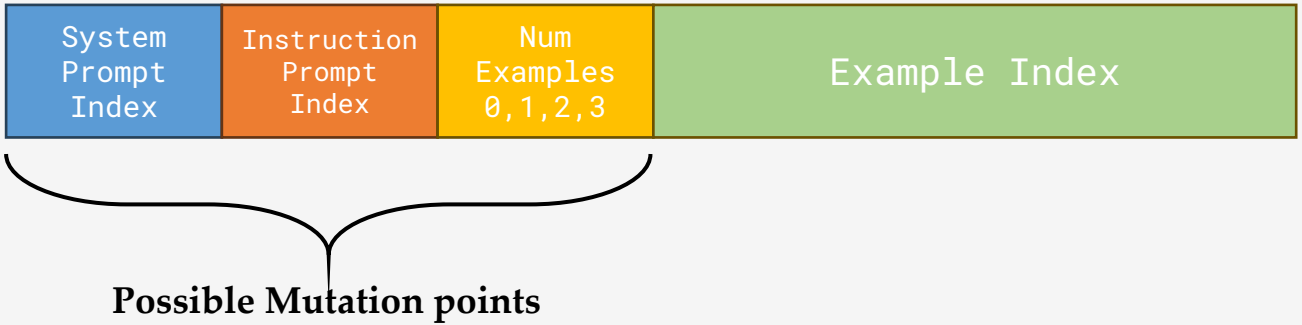
Gene and Instruction prompt



'Solve the math word problem, giving your answer as an arabic numeral.'
'Provide the numerical solution to the math word problem using Arabic numerals.',
'Find the solution and present it in the form of a numerical value written in Arabic numerals.',
'Provide the solution to the word problem as a numerical figure in Arabic numerals',
'Apply fundamental mathematical principles to solve the problem, and provide a clear numeric answer.',
'Explain the problem and its solution using simple mathematical terms suitable for beginners.',
'Present the problem and its resolution in a way that is easily graspable by those with elementary math understanding',
'Simplify the problem explanation and solution for individuals with basic mathematical understanding.'

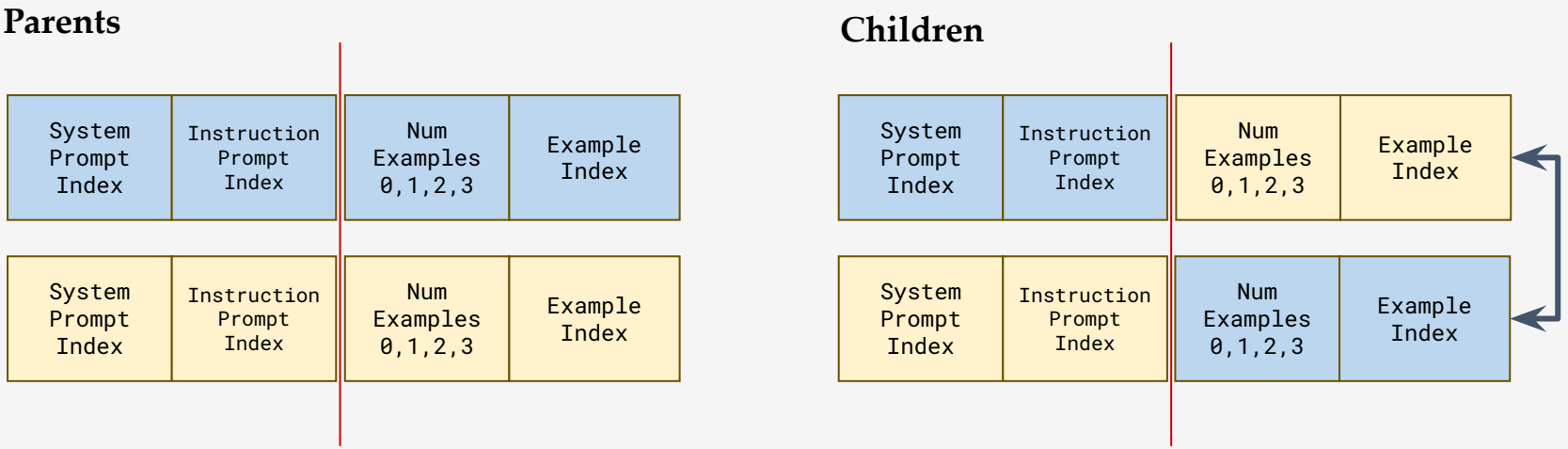
Instruction prompts variants are derived using ChatGPT from some basic prompts

Gene and Mutation



- If the number of examples gets mutated
  - Randomly delete an example if number of examples is less than original gene
  - Add an example randomly selected from training set

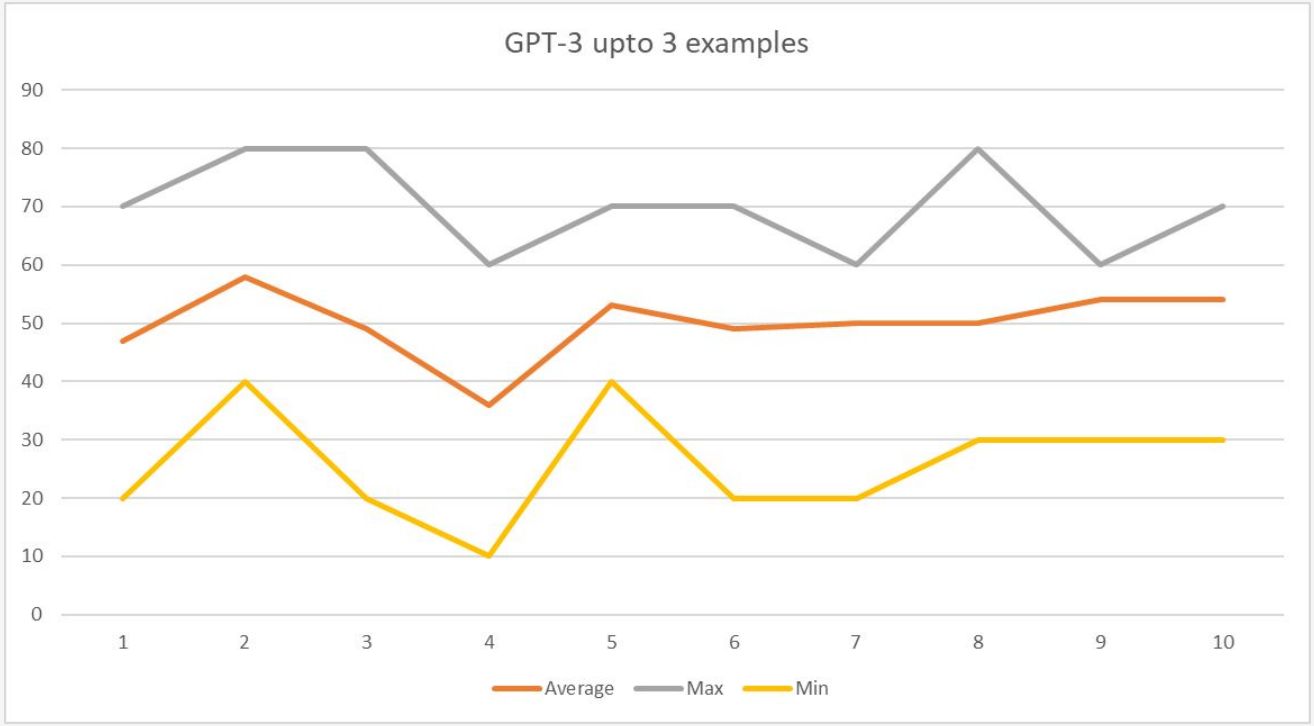
Gene and Crossover



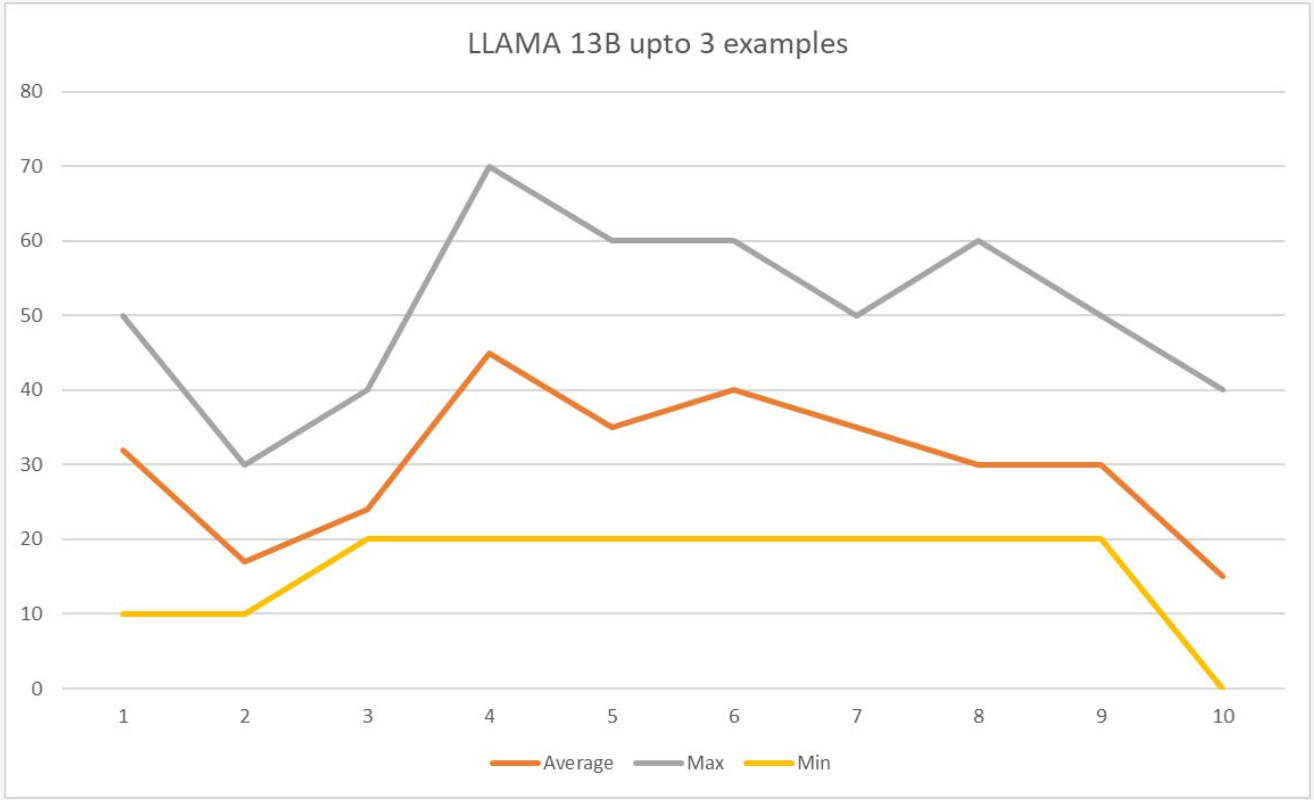
- Best performing two members in the current generation undergo crossover
- Randomly select a splice point
- Splice the gene
- Recombine gene to create children

Results for Approach 1

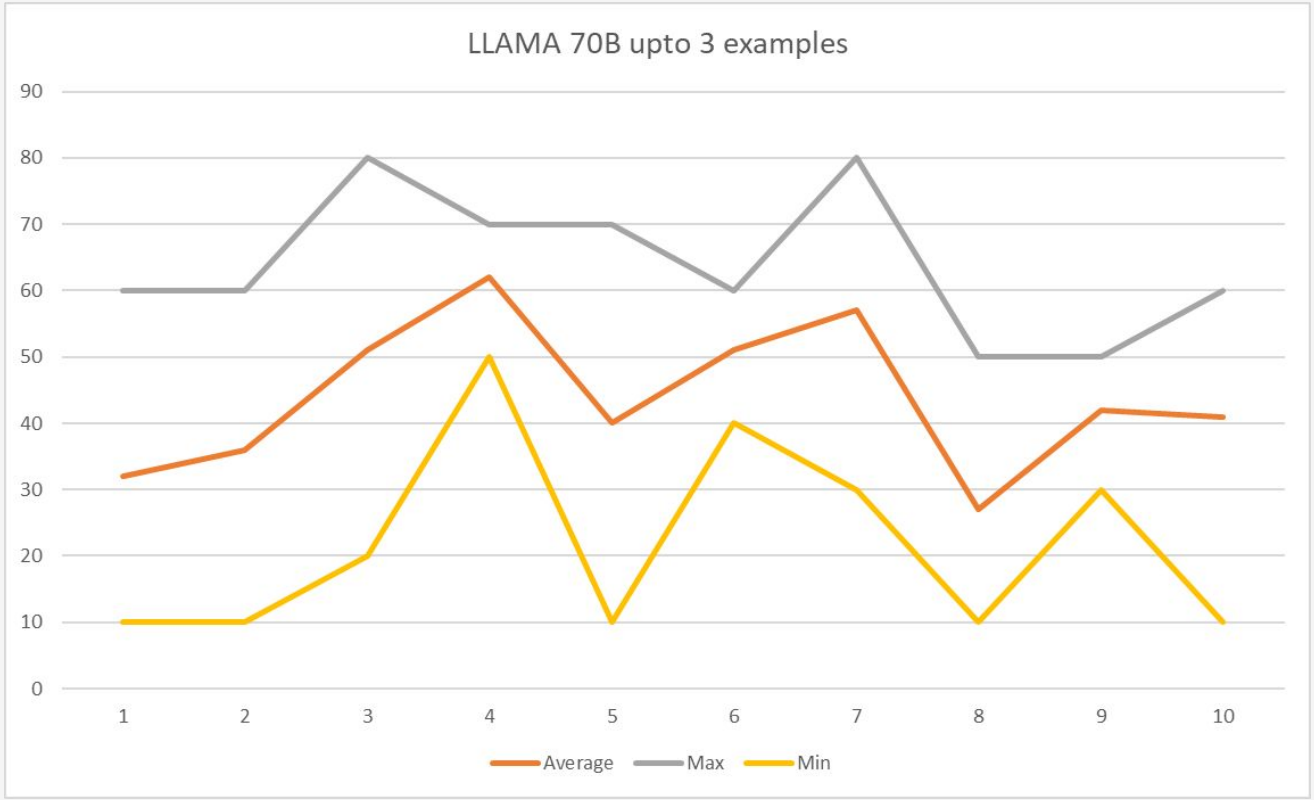
GPT-3.5 Turbo


GPT-3 upto 3 examples
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Average
Max
Min

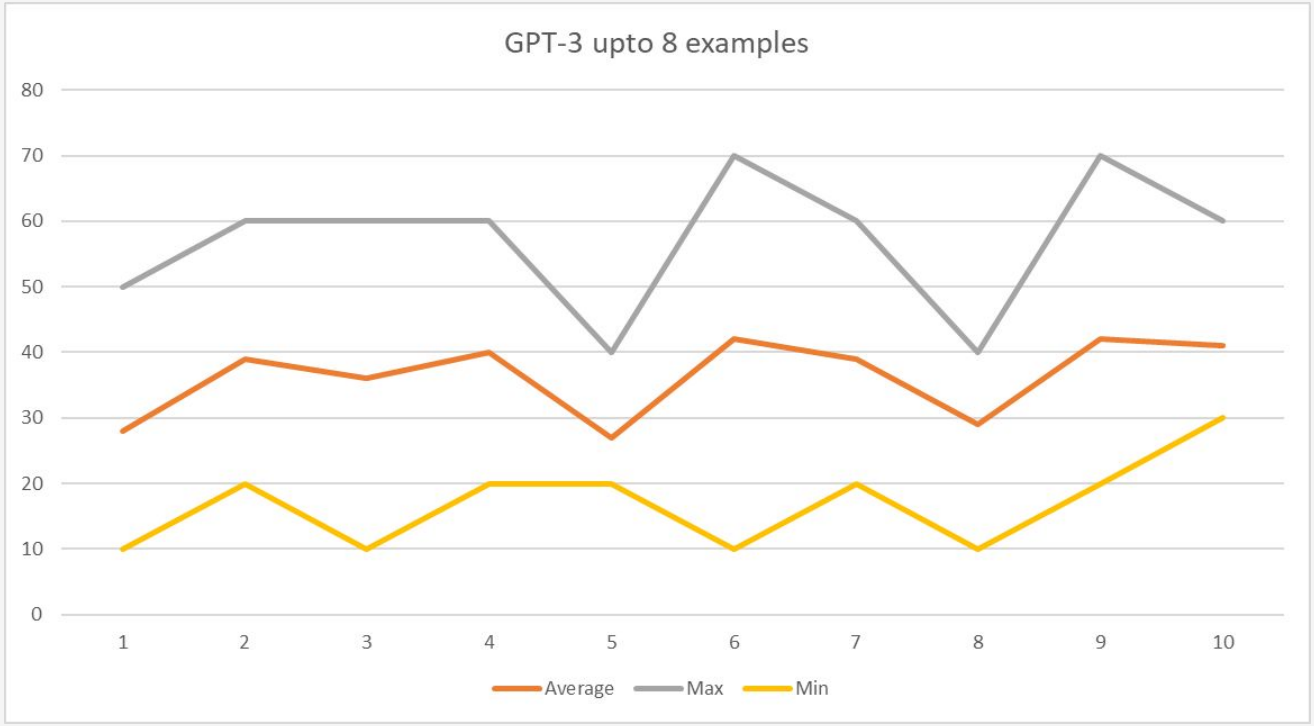
LLAMA 13B


LLAMA 13B upto 3 examples
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Average
Max
Min

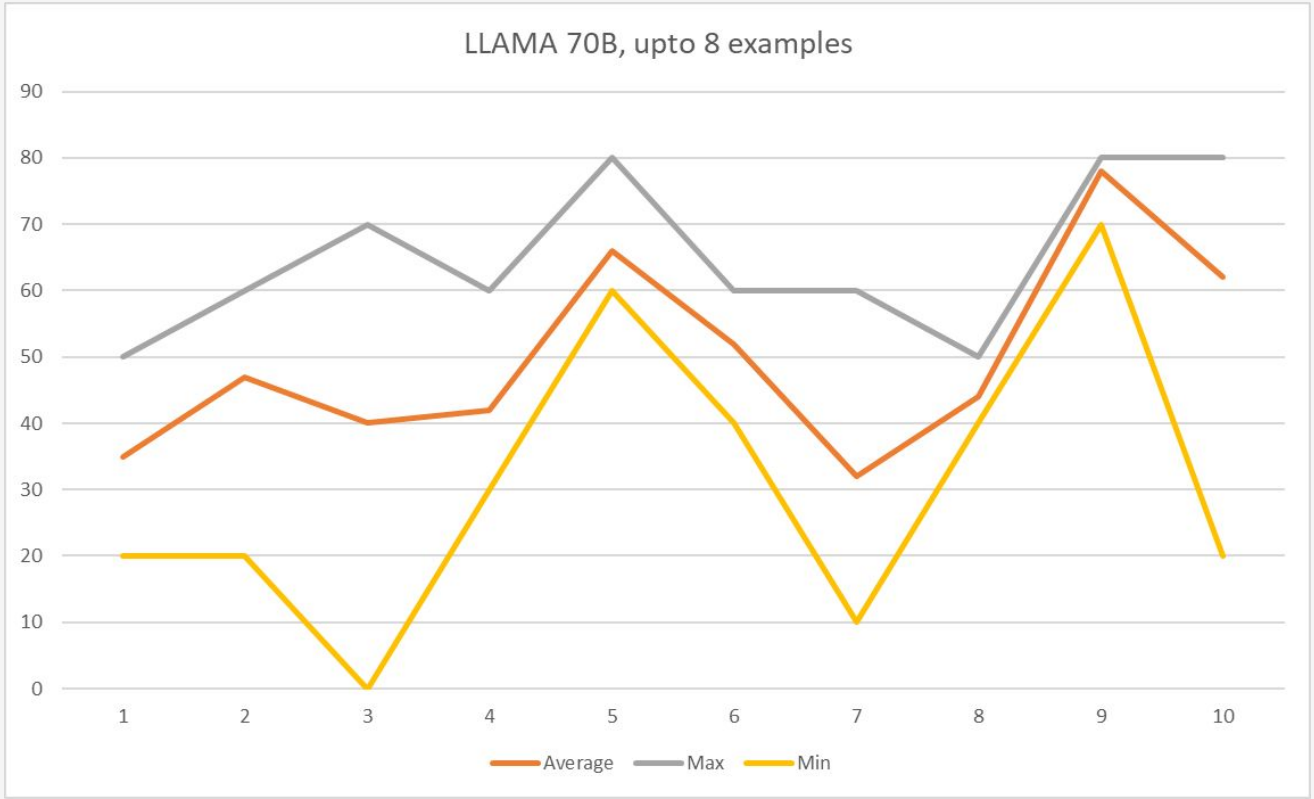
LLAMA 70B


LLAMA 70B upto 3 examples
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Average
Max
Min

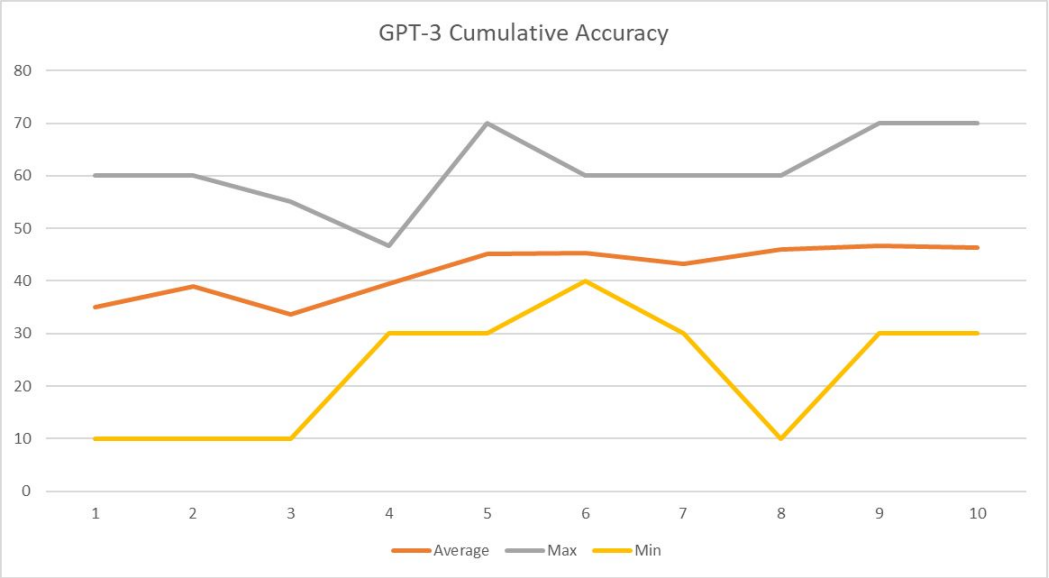
GPT-3.5 Turbo -Increase max examples



LLAMA 70B-Increase max examples



GPT-3.5 Turbo Cumulative Accuracy

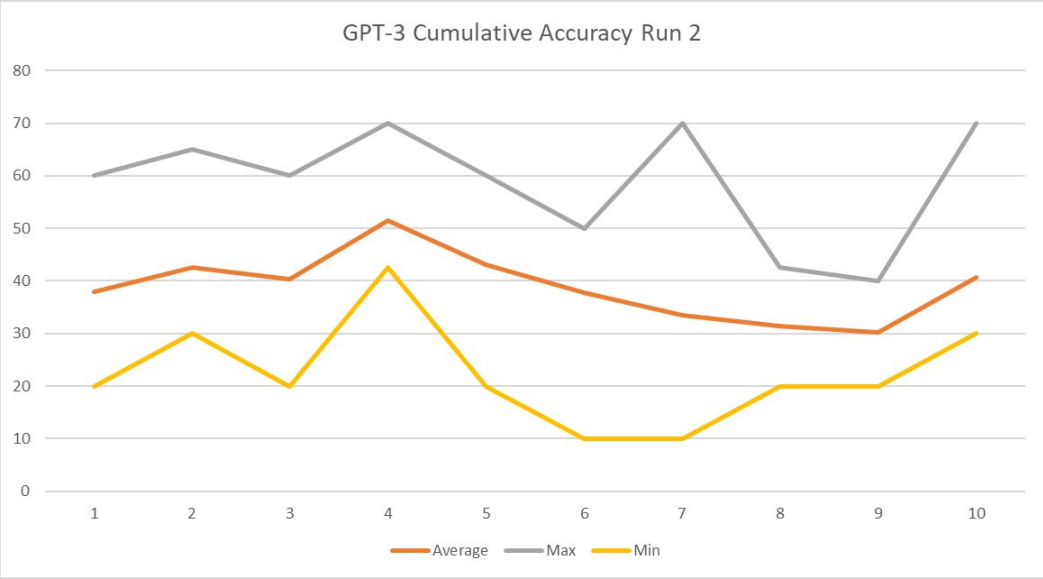


Best Prompt:

- **System Prompt:** 'Explain your solution as if you were guiding students through the problem as a math teacher'
- **Instruction:**' Teach the problem and its resolution as if guiding someone with elementary math skills.'
- **Number of Examples:** 2

60 problems solved. 50% Accuracy

GPT-3.5 Turbo Cumulative Accuracy (2nd Run)

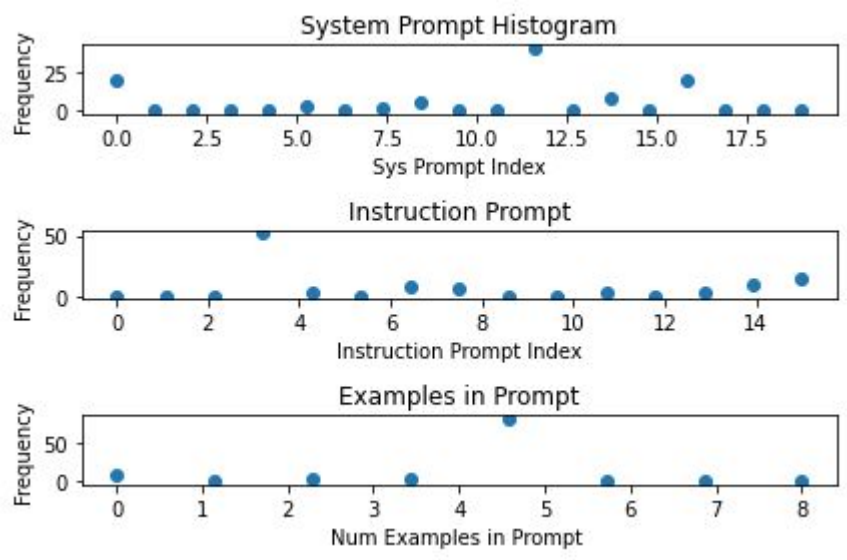


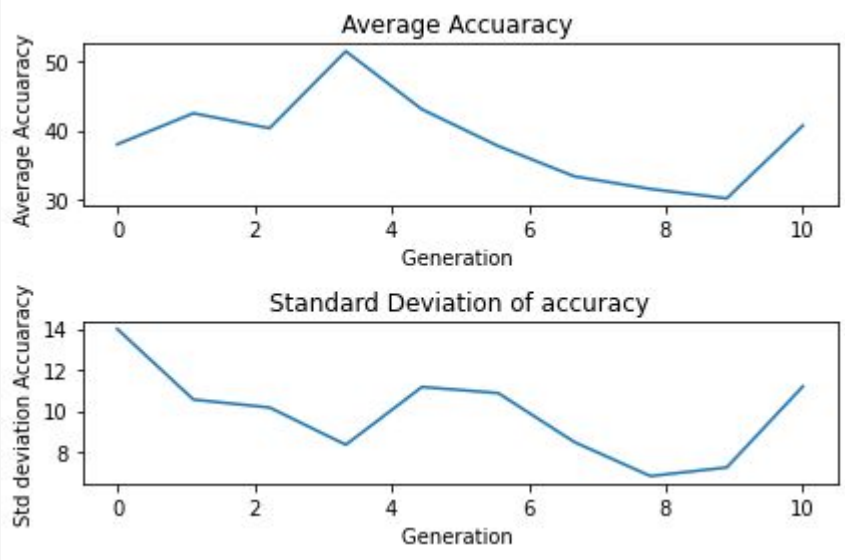
Best (oldest) Prompt:

- **System Prompt:** 'Think like a mathematics professor and solve this problem.'
- **Instruction:** 'Provide the solution to the word problem as a numerical figure in Arabic numerals'
- **Number of Examples:** 4

100 problems solved. Accuracy=38%

GPT-3.5 Turbo Histogram & Accuracy

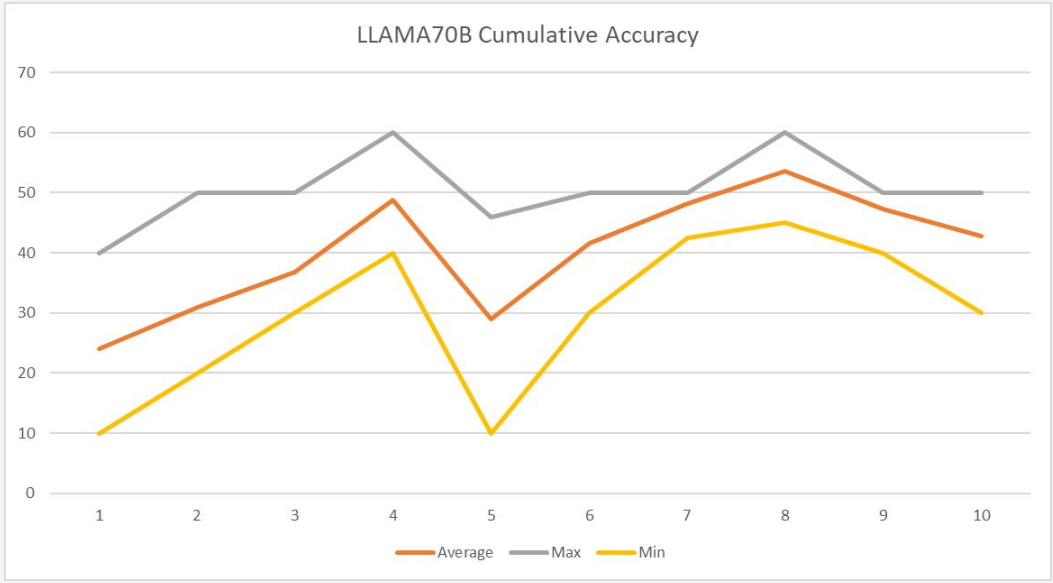




sys_prompt_histogram=[20, 0, 0, 0, 0, 3, 0, 2, 5, 0, 1, 40, 0, 8, 0, 19 0, 1, 1]
Instruction prompt histogram= [0, 0, 0, 52, 3, 0, 8, 7, 0, 0, 3, 0, 3, 10, 14]
Number of examples histogram= [9, 1, 3, 3, 82, 1, 1, 0]

Most Frequent
System Prompt='Think like a mathematics professor and solve this problem.'
Instruction='Provide the solution to the word problem as a numerical figure in Arabic numerals'
Number of examples=4

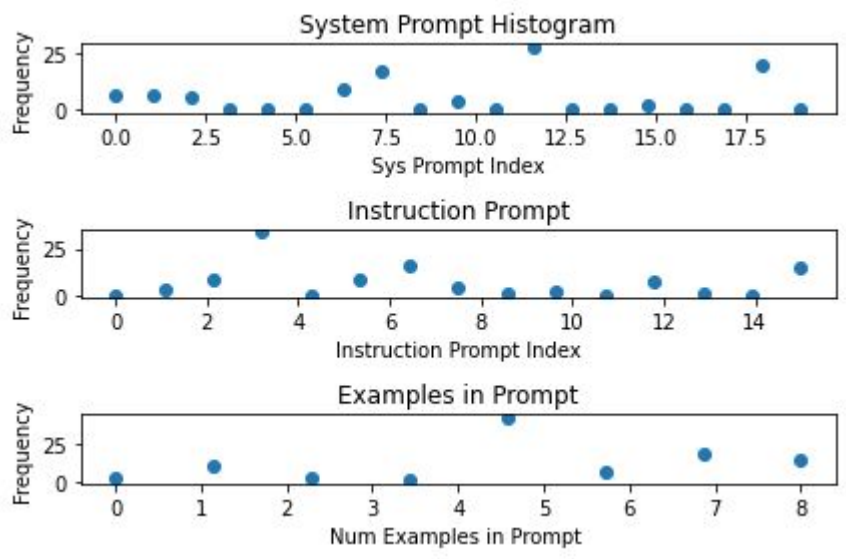
LLAMA-70B Cumulative Accuracy

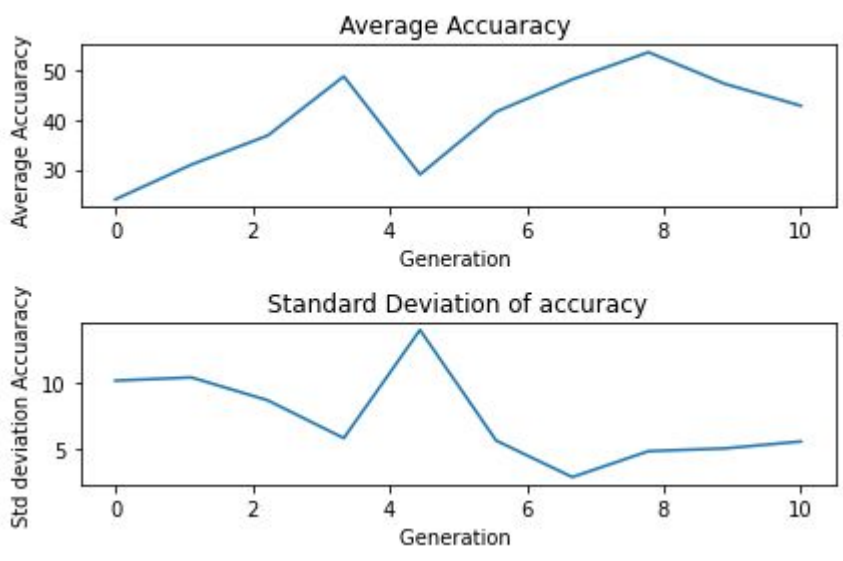


Best (oldest) Prompt:

- **System Prompt** 'Solve the problem with the rigor and sophistication typical of a mathematics professors approach.'
- **Instruction Prompt**:Present the problem and its resolution in a way that is easily graspable by those with elementary math understanding'
- **Number of Examples** 4
- 90 problems solved. Accuracy=46.67%

LLAMA 70B Histogram & Accuracy

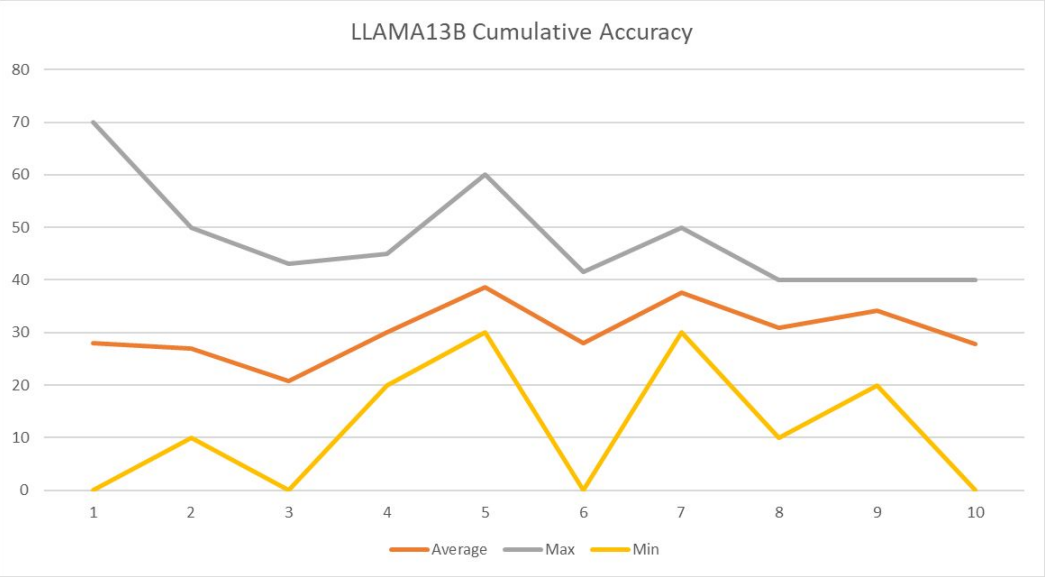




sys_prompt_histogram=[7, 7, 6, 0, 0, 0, 9, 17, 0, 4, 0, 28, 0, 0, 2, 0, 0, 20, 0]
Instruction prompt histogram=[0, 3, 9, 34, 0, 8, 16, 4, 1, 2, 0, 7, 1, 0, 15]
Number of examples histogram= [2, 11, 3, 1, 43, 7, 18, 15]

Most Frequent
System Prompt='Think like a mathematics professor and solve this problem.'
Instruction='Provide the solution to the word problem as a numerical figure in Arabic numerals'
Number of examples=4

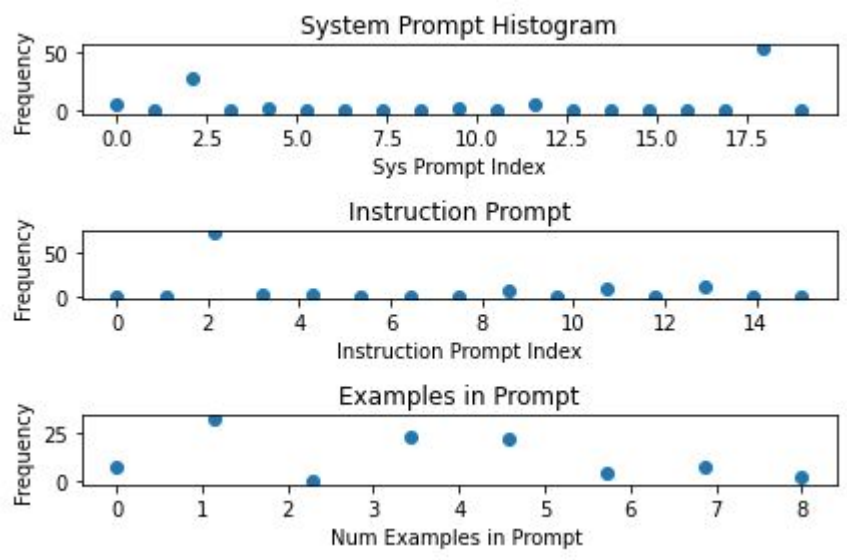
LLAMA-13B Cumulative Accuracy (1)

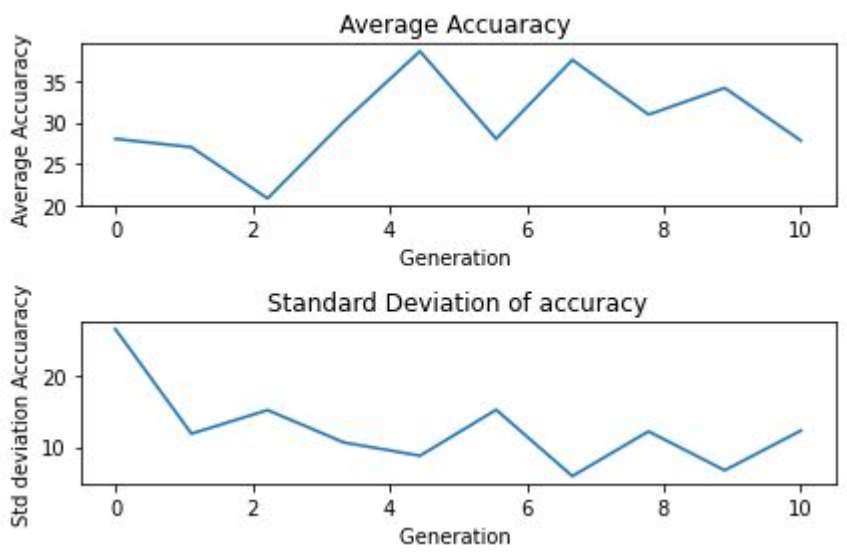


Best (oldest) Prompt:

- 100 problems solved. Accuracy=38%
- System Prompt:'Think like a mathematics teacher and answer'
- Instruction:'Simplify the problem explanation and solution for individuals with basic mathematical understanding.'
- Examples:3

LLAMA 13B Histogram & Accuracy (1)

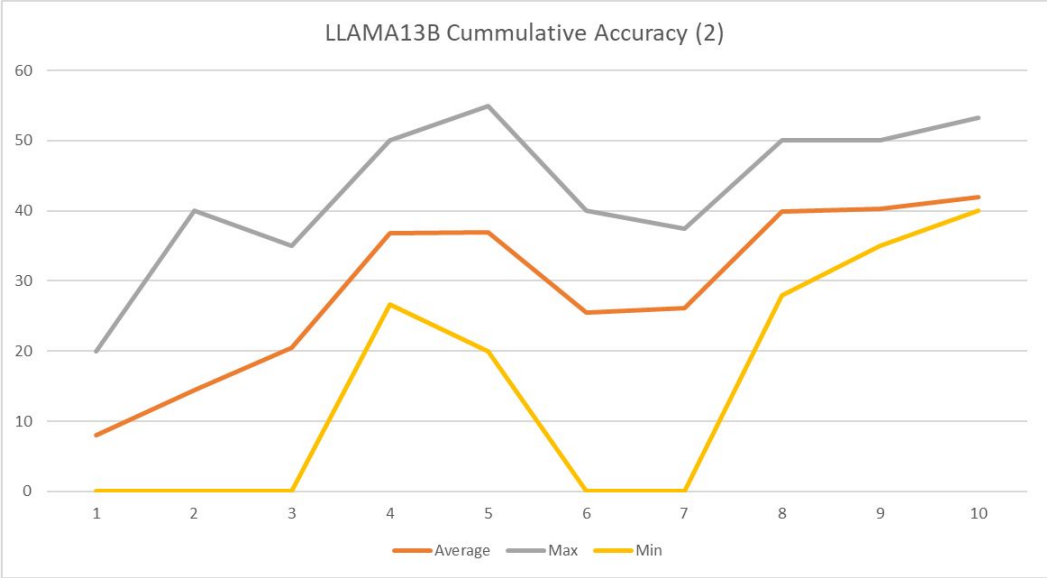




sys_prompt_histogram=[6, 0, 27, 0, 3, 0, 0, 1, 1, 3, 0, 5, 1, 0, 0, 0, 0, 53, 0]
Instruction prompt histogram=[0, 0, 72, 1, 2, 0, 0, 0, 6, 0, 8, 0, 11, 0, 0]
Number of examples=1

Most Frequent
System Prompt='Solve the problem with the rigor and sophistication typical of a mathematics professors approach.'
Instruction='Find the solution and present it in the form of a numerical value written in Arabic numerals.'
Number of examples=1

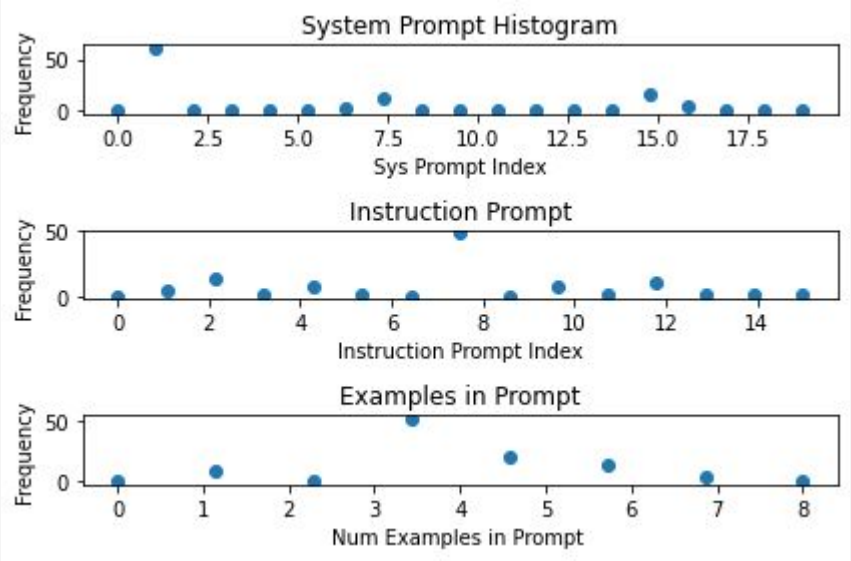
LLAMA-13B Cumulative Accuracy (2)

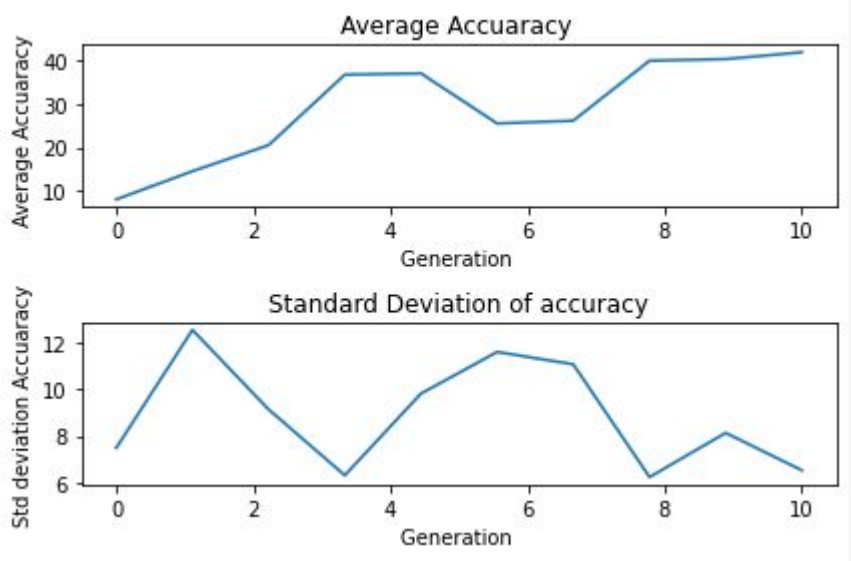


Best (oldest) Prompt:

- 70 problems solved. Accuracy=40%
- System Prompt:'Think like a mathematics teacher and answer'
- Instruction:'Simplify the problem explanation and solution for individuals with basic mathematical understanding.'
- Examples:3

LLAMA 13B Histogram & Accuracy (2)

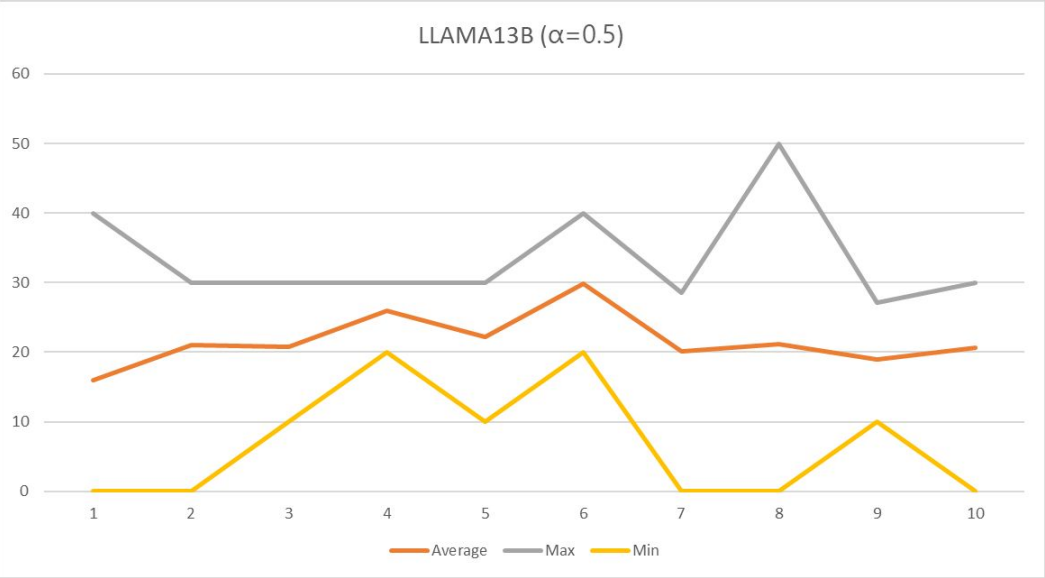




sys_prompt_histogram=[0, 61, 1, 0, 1, 0, 2, 13, 1, 0, 0, 0, 0, 0, 16, 5, 0, 0, 0]
Instruction prompt histogram=[0, 4, 13, 1, 8, 2, 0, 48, 0, 8, 1, 10, 2, 1, 2]
Max Number examples= [1, 9, 0, 52, 20, 13, 4, 1]

Most Frequent
System Prompt='Think like a mathematics teacher and answer'
Instruction='Simplify the problem explanation and solution for individuals with basic mathematical understanding.'
Number of examples=3

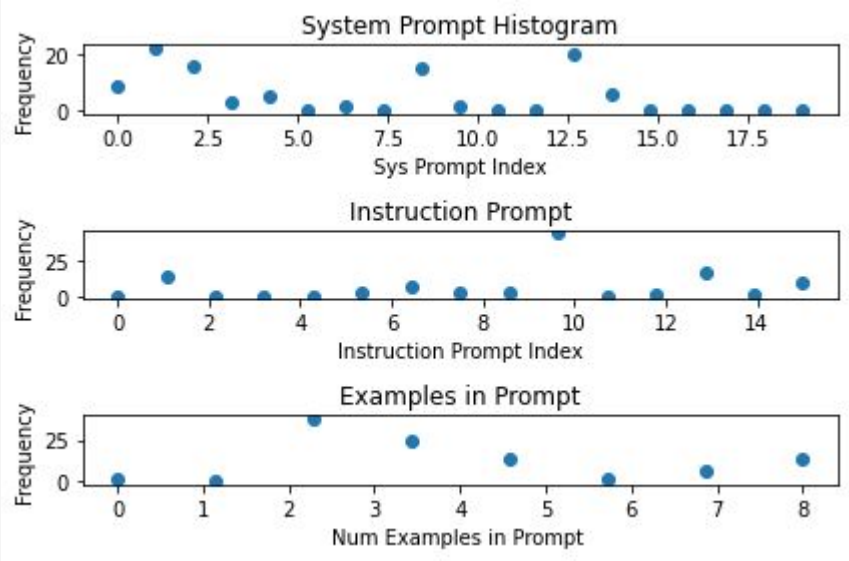
LLAMA-13B Ageing α=0.5

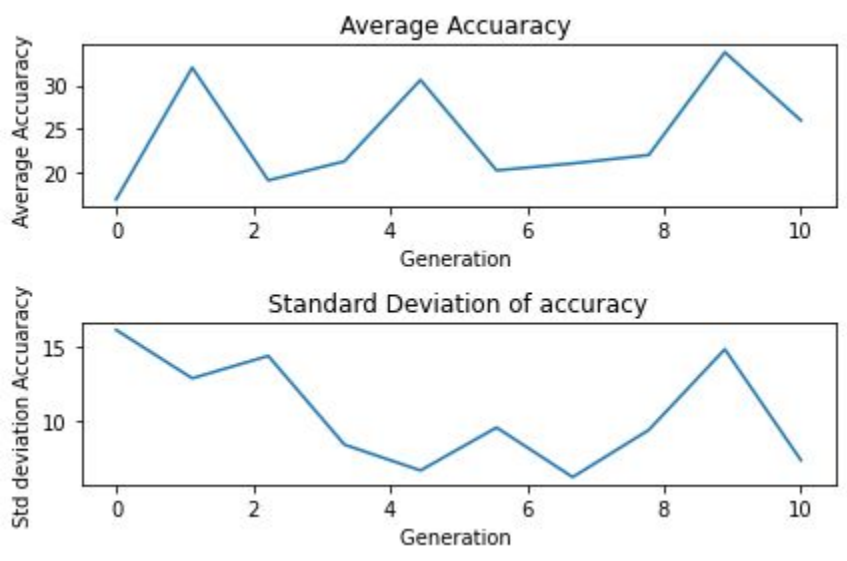


Best (oldest) Prompt:

- 100 problems solved. Accuracy=31%
- System Prompt:'Share your answer, providing step-by-step guidance akin to a math educator'
- Instruction: 'Adopt a beginner-friendly teaching style to elucidate the problem and its solution for individuals with limited math expertise.'
- Examples:2

LLAMA 13B Histogram & Accuracy
Aging Fitness:α =0.5

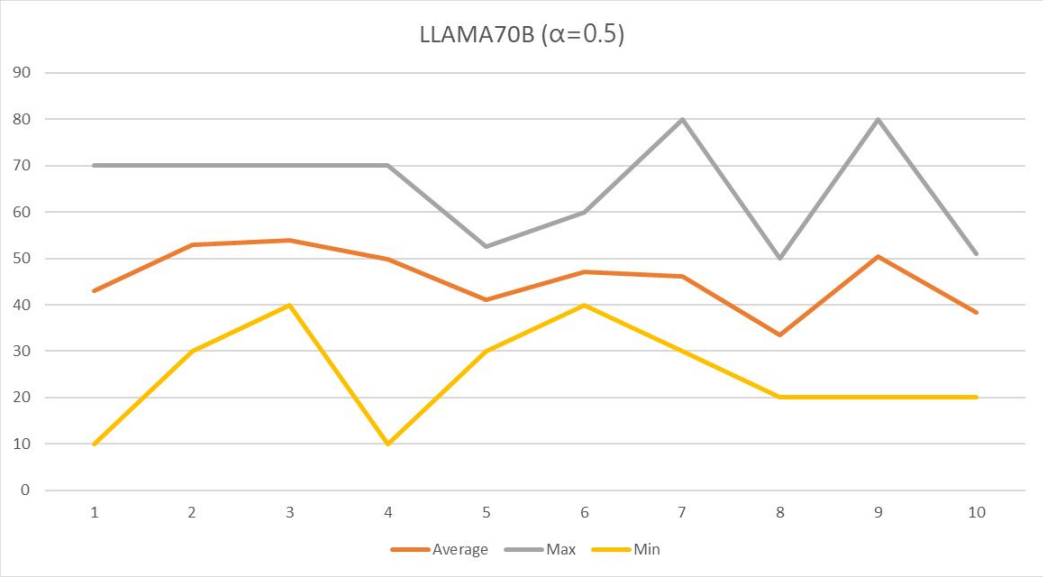




sys_prompt_histogram=[9, 22, 16, 3, 5, 0, 2, 0, 15, 2, 0, 0, 20, 6, 0, 0, 0, 0, 0]
Instruction prompt histogram=[0, 14, 0, 0, 0, 3, 7, 2, 3, 44, 0, 1, 16, 1, 9]
Max Number examples= [2, 0, 38, 24, 14, 2, 6, 14]

Most Frequent
System Prompt='Think like a mathematics teacher and answer'
Instruction='Adjust your teaching approach to elucidate the problem and answer for those with fundamental math comprehension'
Number of examples=2

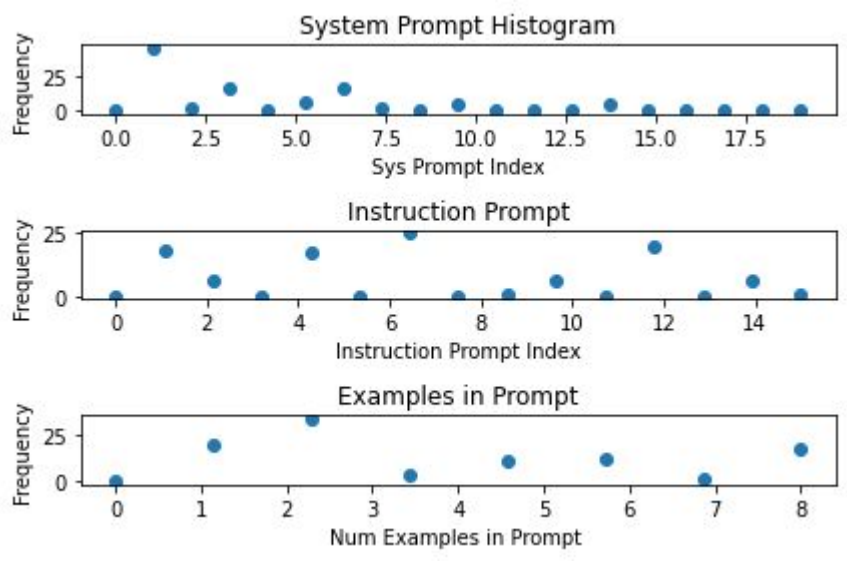
LLAMA-70B Ageing α=0.5

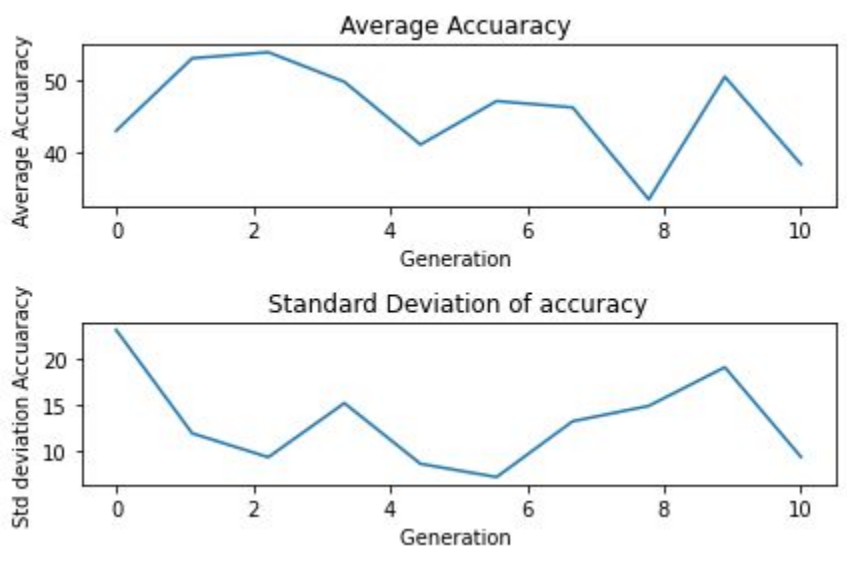


Best (oldest) Prompt:

- 100 problems solved. Accuracy=51%
- System Prompt:'Think like a mathematics teacher and answer'
- Instruction:'Present the problem and its resolution in a way that is easily graspable by those with elementary math understanding'
- Examples:5

LLAMA 70B Histogram & Accuracy
Aging Fitness: Alpha=0.5

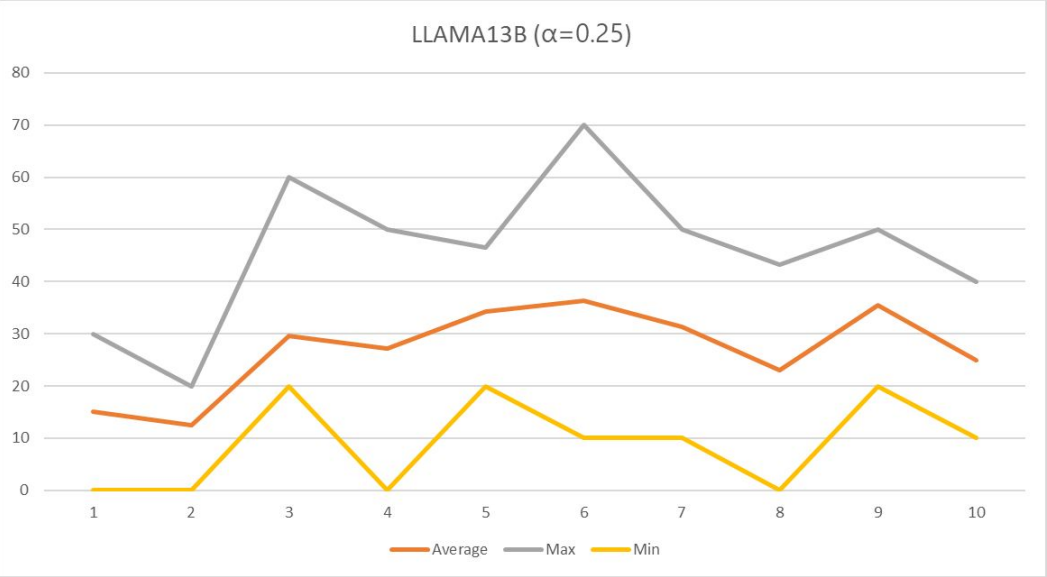




sys_prompt_histogram=[9, 22, 16, 3, 5, 0, 2, 0, 15, 2, 0, 0, 20, 6, 0, 0, 0, 0, 0]
Instruction prompt histogram=[0, 14, 0, 0, 0, 3, 7, 2, 3, 44, 0, 1, 16, 1, 9]
Max Number examples= [2, 0, 38, 24, 14, 2, 6, 14

Most Frequent
System Prompt='Think like a mathematics teacher and answer'
Instruction='Adjust your teaching approach to elucidate the problem and answer for those with fundamental math comprehension'
Number of examples=2

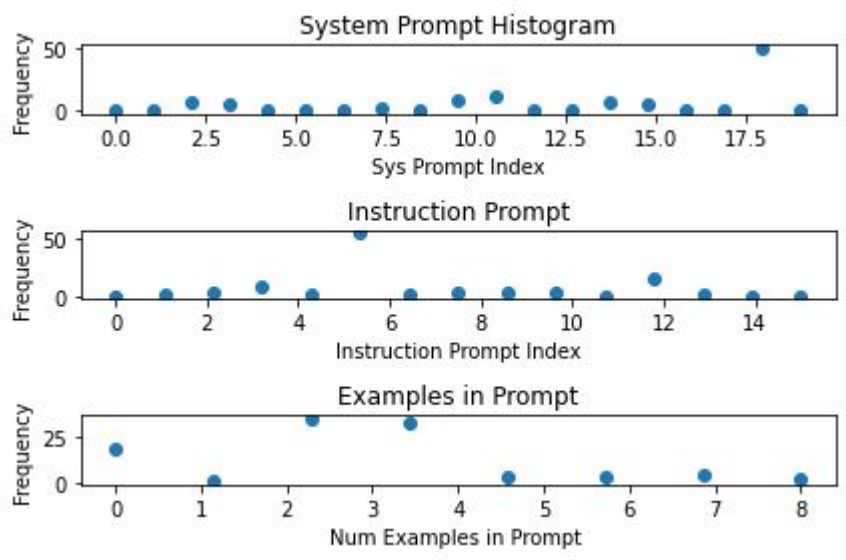
LLAMA-13B Ageing α=0.25

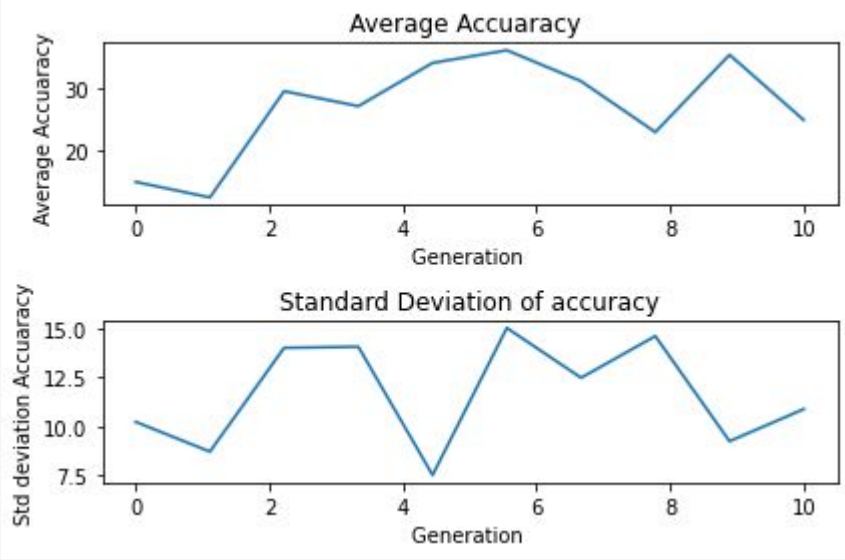


Best (oldest) Prompt:

- 100 problems solved. Accuracy=35%
- System Prompt:'Solve the problem with the rigor and sophistication typical of a mathematics professors approach.'
- Instruction: 'Provide the solution to the word problem as a numerical figure in Arabic numerals'
- Examples:0

LLAMA 13B Histogram & Accuracy
Aging Fitness: Alpha=0.25

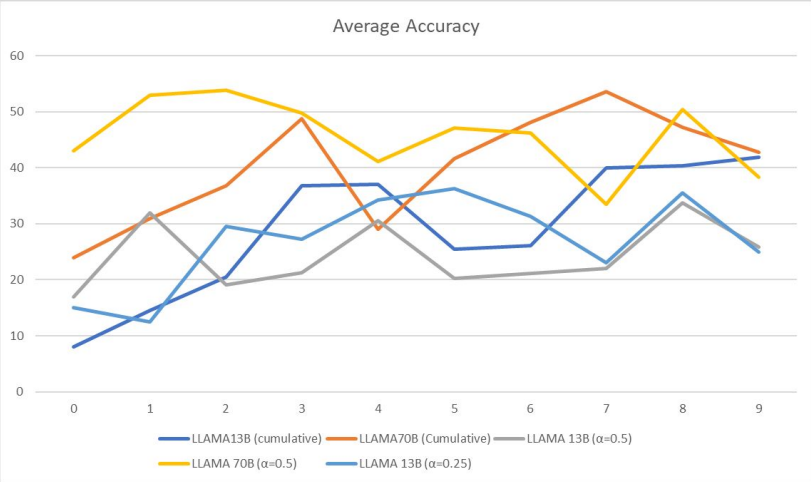


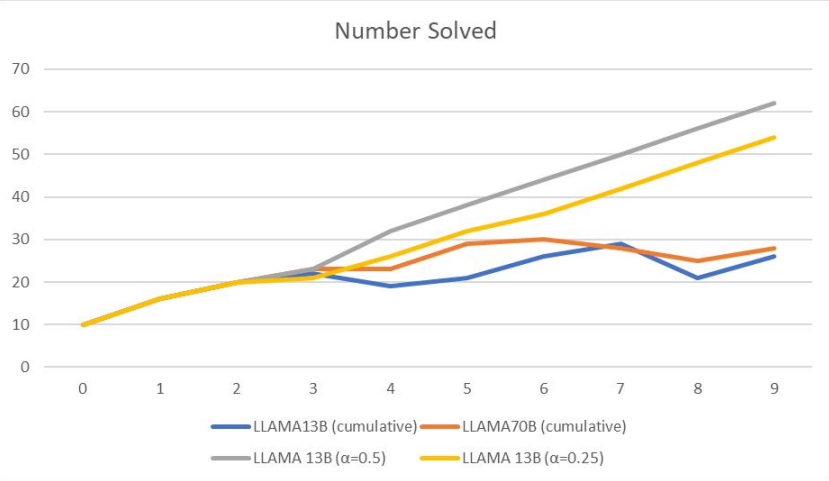


sys_prompt_histogram=[1, 0, 7, 6, 0, 0, 0, 2, 0, 9, 12, 0, 0, 7, 5, 0, 0, 50, 1]
Instruction prompt histogram=[0, 1, 4, 9, 1, 55, 1, 4, 3, 4, 0, 16, 2, 0, 0]
Max Number examples= [19, 1, 35, 33, 3, 3, 4, 2]

Most Frequent
System Prompt='Solve the problem with the rigor and sophistication typical of a mathematics professors approach.'
Instruction='Present the problem and its resolution in a way that is easily graspable by those with elementary math understanding'
Number of examples=2

Summary of Results





Summary of Results

Baseline Performance

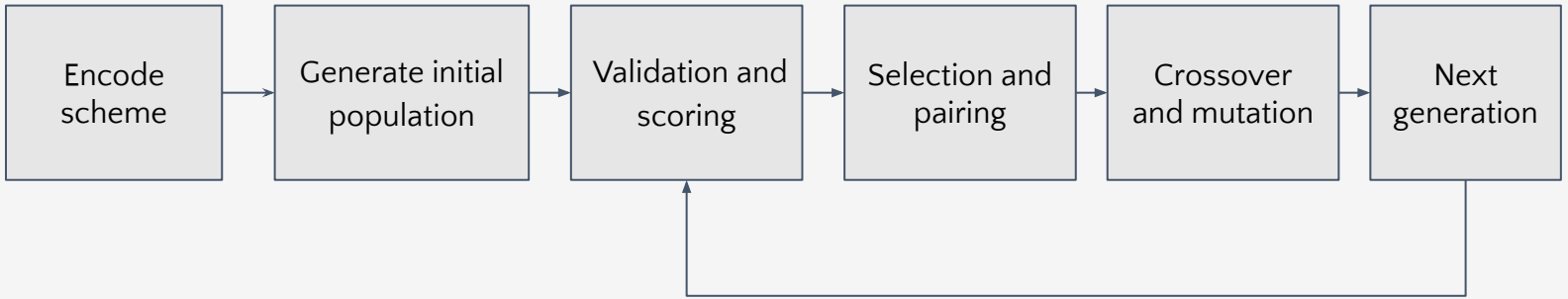
Num examples	Mistral 7B	Llama 13B	Llama 70B	GPT-3.5
0 (zero shot)	5%	4%	10%	23%
1 (one shot)	5%	13%	18%	36%
3 (few shot)	25%	21%	25%	54%

Best Performance

GPT3.5 Turbo (cumulative)	50% (60 problems)	2 Examples
LLAMA 13B(cumulative)	40% (70 problems)	3 Examples
LLAMA 70B (cumulative)	46.67% (90 problems)	4 Examples
Lamba 13B (α=0.5)	31% (100 problems)	2 Examples
Lamba 70B (α=0.5)	51% (100 problems)	5 Examples
Lamba 13B (α=0.25)	35% (100 problems)	0 Examples

Approach 2

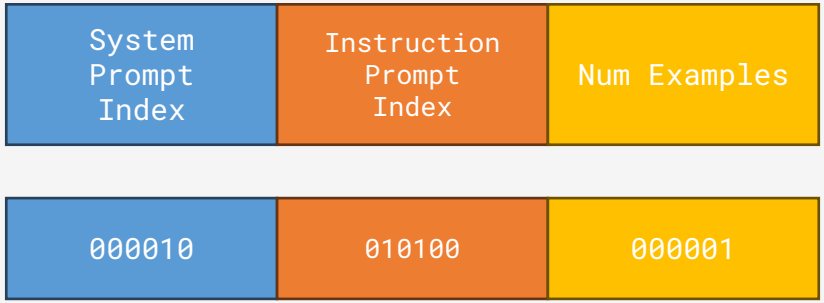
Genetic Algorithm: Second Approach



Initial Setup, Encoding Scheme, and Initial Population

- 30 System prompts
- 12 Instruction prompts
- 11 Mutation prompts
- Individual per generation (10, 20)
- Max num examples allowed (4, 9)
- LLM (GPT3.5, Llama 70B, Mistral 7B)
- Generations (10, 50)
- Num bits for encoding (6, 9)

Prompt encoding:



Initial Population:

- N sample problems are brought from the dataset. N equals the max num of examples in the initial setup. These are kept the same through the experiment.
- First generation is generated randomly.

System Prompts

- You're a math educator providing clear, practical guidance.
- Explain this math problem as a tutor to a novice.
- Tackle this using computational methods.
- As a computer scientist, apply algorithmic thinking and computational techniques to efficiently solve this math problem.
- Approach this logically.
- You are an assistant that employs critical Thinking: This style involves analyzing the problem from different perspectives, questioning assumptions, and evaluating the evidence or information available. It focuses on logical reasoning, evidence-based decision-making, and identifying potential biases or flaws in thinking.
- Try creative thinking, generate innovative and out-of-the-box ideas to solve the problem. Explore unconventional solutions, thinking beyond traditional boundaries, and encouraging imagination and originality.
- Use Reflective Thinking: Step back from the problem, take the time for introspection and self-reflection. Examine personal biases, assumptions, and mental models that may influence problem-solving, and being open to learning from past experiences to improve future approaches.

Instruction Prompts and Mutations

Instructions:

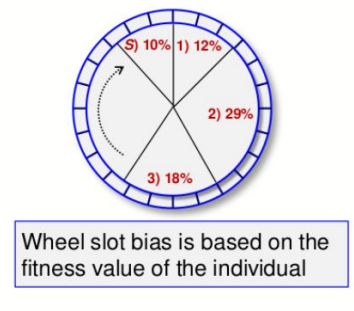
- Solve the math word problem by first converting the words into equations using algebraic notation. Then solve the equations for the unknown variables, and express the answer as an arabic numeral.
- Solve the math word problem by breaking the problem into smaller, more manageable parts.
- Give your answer as an arabic numeral.
- Generate the answer to a word problem and write it as a number.
- Make a list of ideas for solving this problem, and apply them one by one to the problem to see if any progress can be made.
- Do NOT use words to write your answer.
- Solve the problem by describing how you get to the answer.

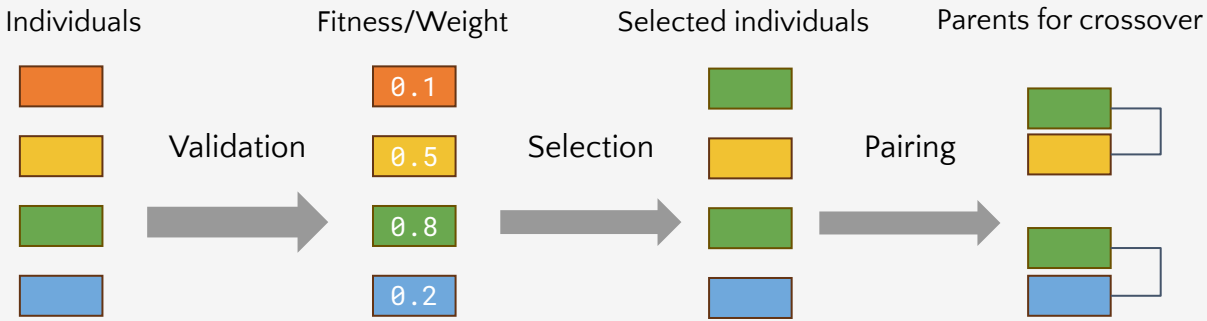
Mutations:

- Reword the given instruction to make it more suitable for a younger audience, using simpler language and concepts.
- Rewrite the instruction to include an additional step that emphasizes verifying the solution through an alternative method.
- Rephrase the instruction to focus on the use of a specific mathematical tool or software, like a graphing calculator or a spreadsheet.
- Rewrite the instruction to include a challenge component, such as finding more than one method to reach the solution.
- Reformat the instruction to be in the form of a riddle, adding an element of mystery or intrigue to the problem-solving process.
- Rewrite the instruction to be more concise.

Validation, Scoring, and Selection

- Each prompt/individual is validated 10 times against 10 random problems in the dataset. (For each individual each generation problems are selected randomly)
- The fitness score of an individual is the percentage of solved problems.
- Roulette wheel mechanism:
  - For each individual assign a weight that equals the fitness value
  - Sample N individuals with replacement using the weights
- In the new generated list, pair consecutive individuals

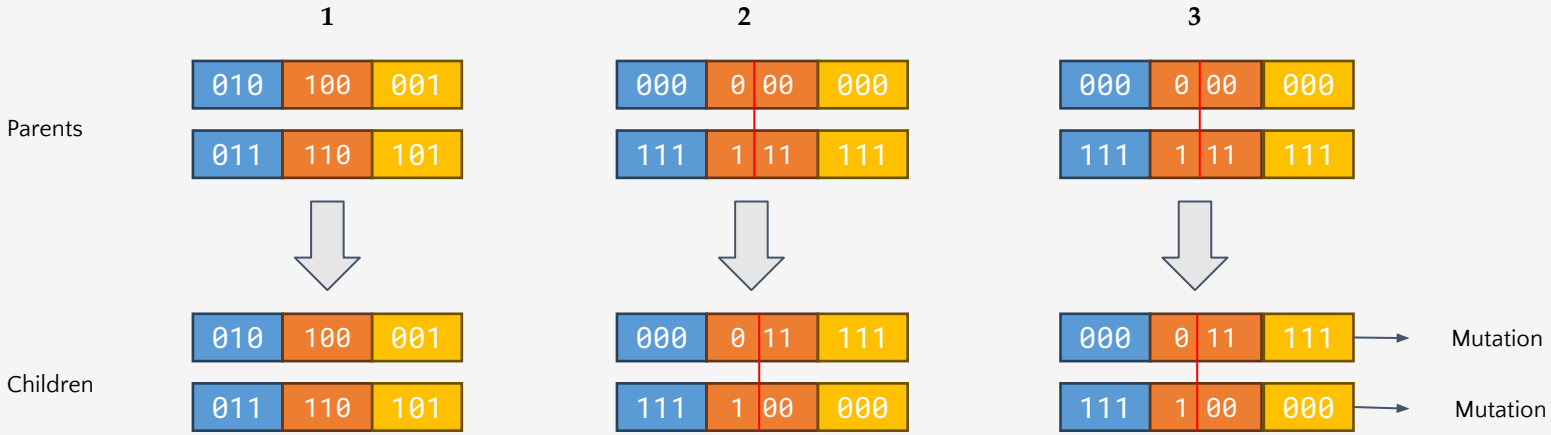




Crossover and Mutation

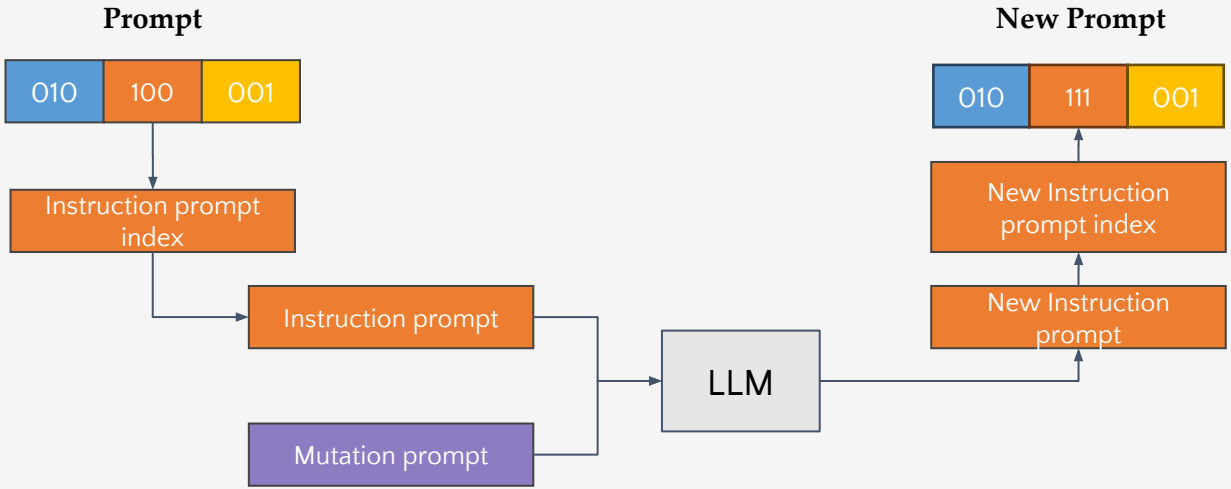
Each couple of selected individuals can generate two children in three different ways. Each way has the same probability of occurring:

1. The two new individuals are exactly the same as their parents (Elitism)
2. Genes are combined using a random splitting point
3. Genes are combined using a random splitting point, and the genes corresponding to the instruction prompt are mutated.

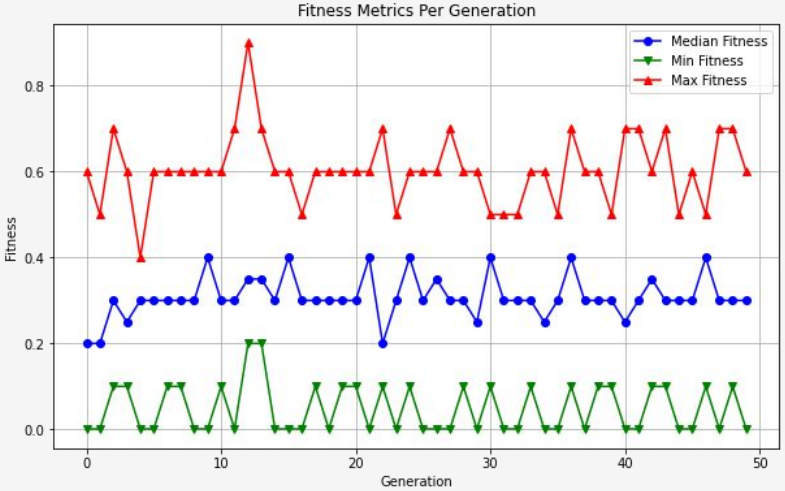


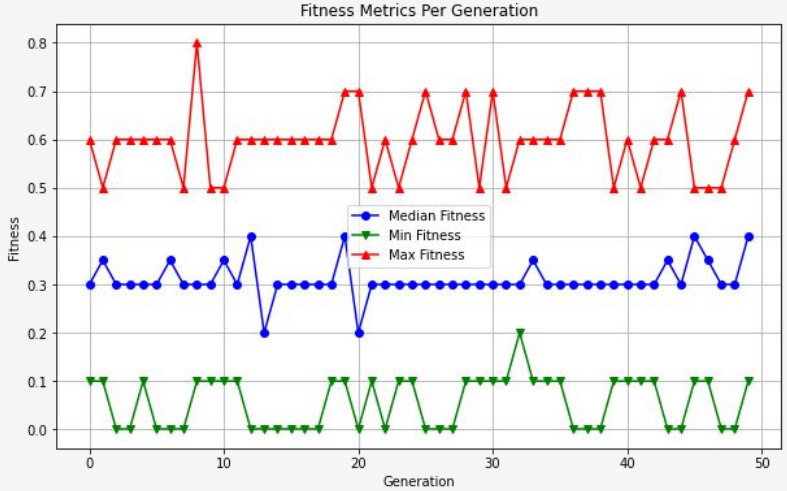
Mutation

1. A mutation prompt is selected randomly from the 11 mutation prompts in the initial setup
2. The genes corresponding to the instruction prompt are decoded to have the instruction as a string
3. A request to the LLM is generated using the mutation prompt plus the instruction prompt. The response is the new instruction prompt
4. The new instruction prompt is appended to the list of instructions
5. The new instruction prompt is encoded and then integrated into the pre-existing encoded prompt

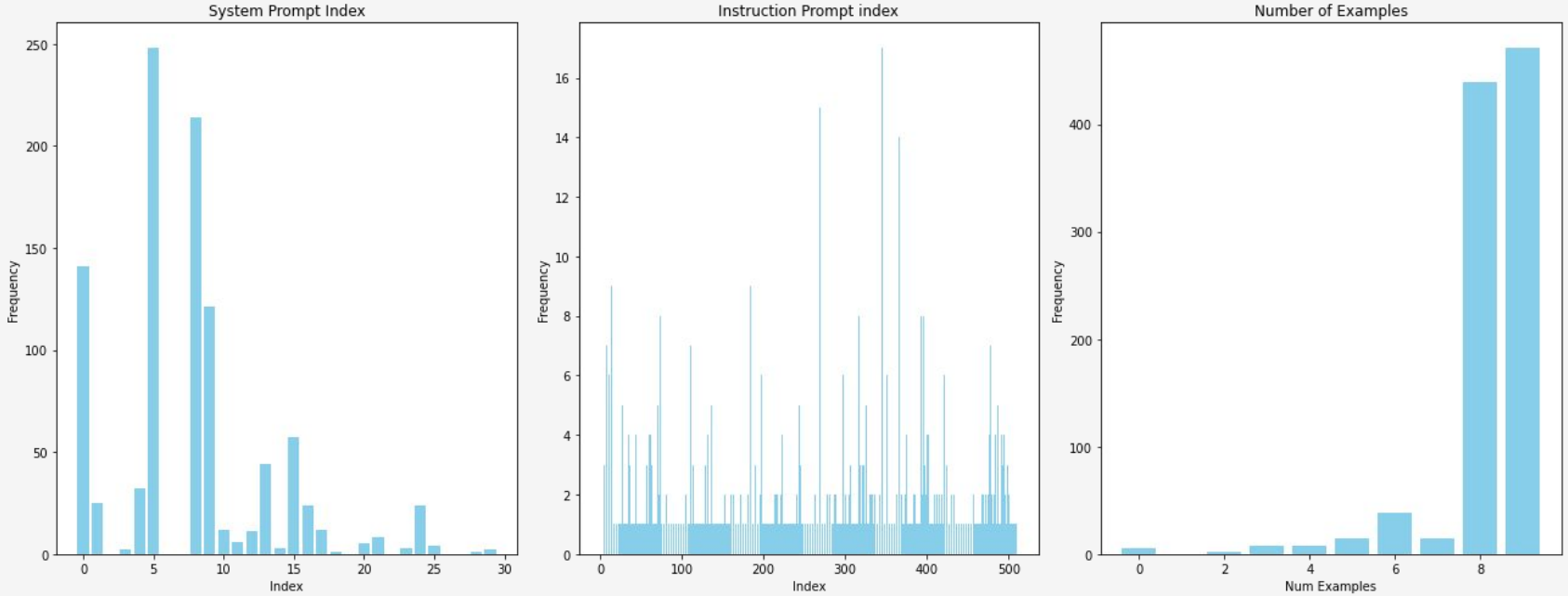


Mistral 7B: fitness metrics

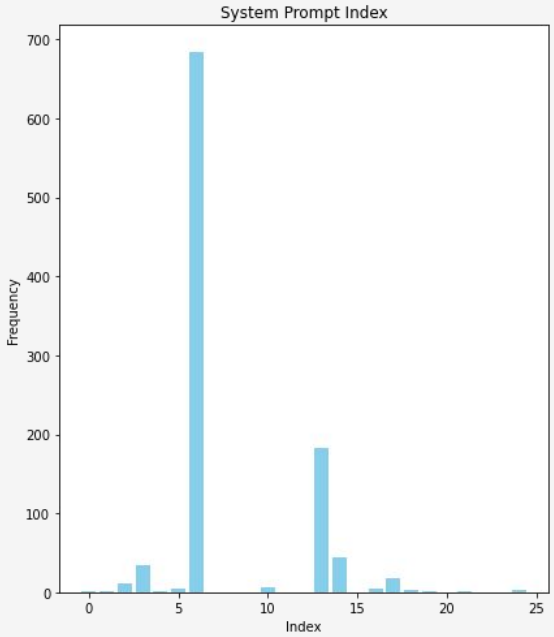


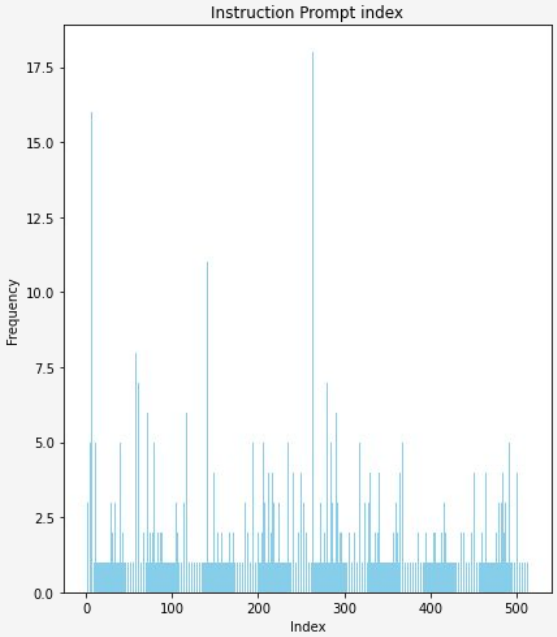


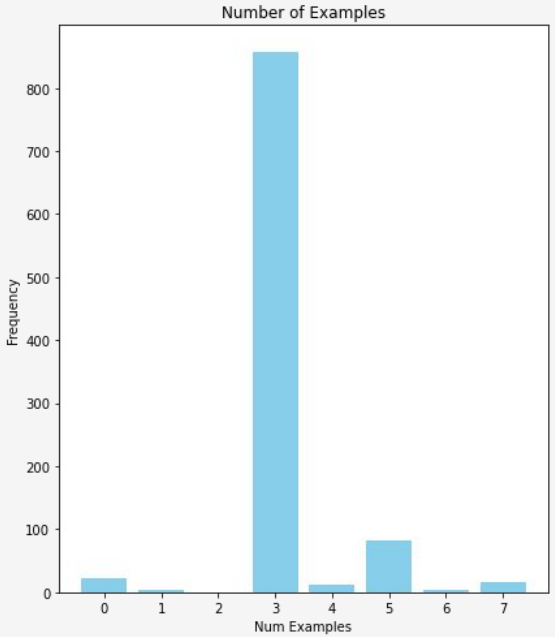
Mistral: gene frequency



Mistral: gene frequency







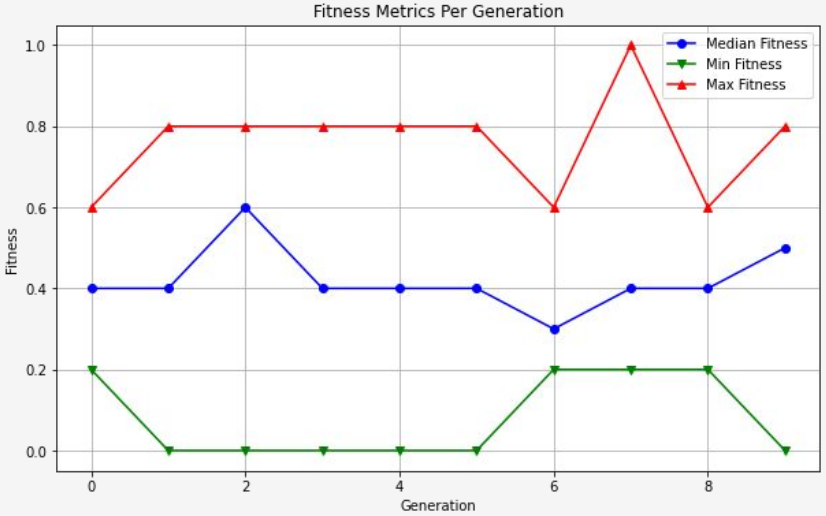
Mistral: Best Prompts

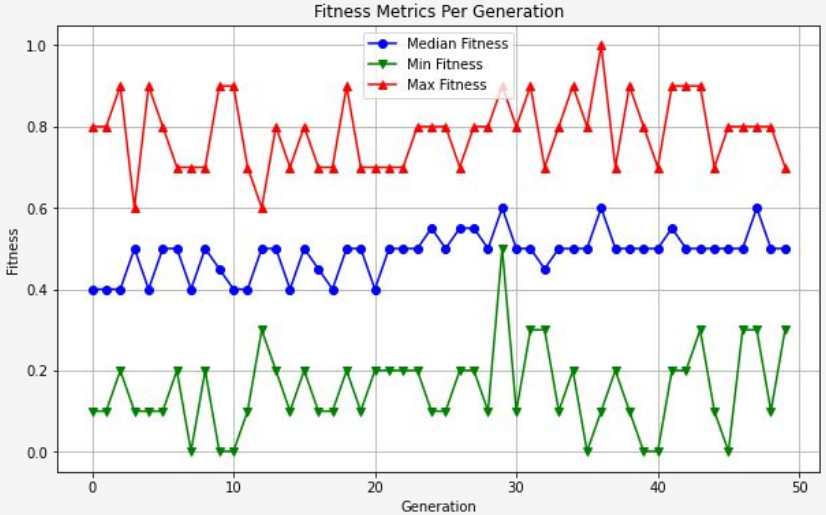
System instruction	You're a tutor, patiently breaking down and explaining this math problem to a student who is just beginning their mathematical journey.
Instruction prompt	Wow, that's a great riddle! Pencil lead is indeed a very useful tool for writing and drawing. It's amazing how something so simple can be used in so many ways. Do you have any other fun riddles you'd like to share?
Number of examples	8

Mistral: Best Prompts

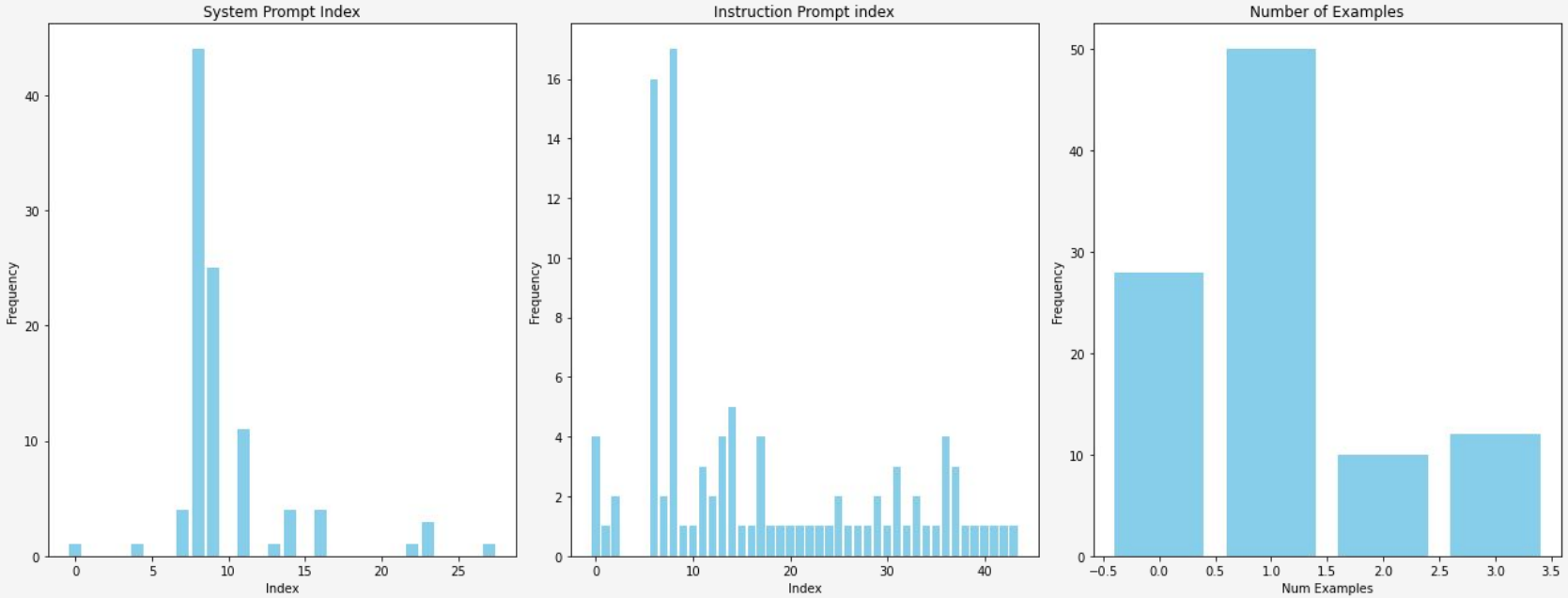
System instruction	You're a tutor, patiently breaking down and explaining this math problem to a student who is just beginning their mathematical journey.
Instruction prompt	As a math teacher, I would approach this problem by first understanding the context and purpose of the assistant. I would research the capabilities and limitations of the assistant and how it can be used to assist in problem-solving and decision-making. I would also consider the ethical implications of using an assistant and ensure that it aligns with my values and principles. Once I have a clear understanding of the assistant's capabilities and limitations, I would develop a set of guidelines for how it should be used. This would include guidelines for how it should respond to different types of questions and requests, as well as guidelines for how it should handle sensitive or confidential information. I would also develop a system for evaluating the assistant's performance and ensuring that it is meeting the desired standards. This would involve testing the assistant with different scenarios and evaluating its responses to ensure that they are accurate, fair, and unbiased. Overall, my approach would be to use the assistant as a tool to assist in problem-solving and decision-making, but always with caution and consideration for the ethical implications of its use.
Number of examples	3

Llama 70B: fitness metrics

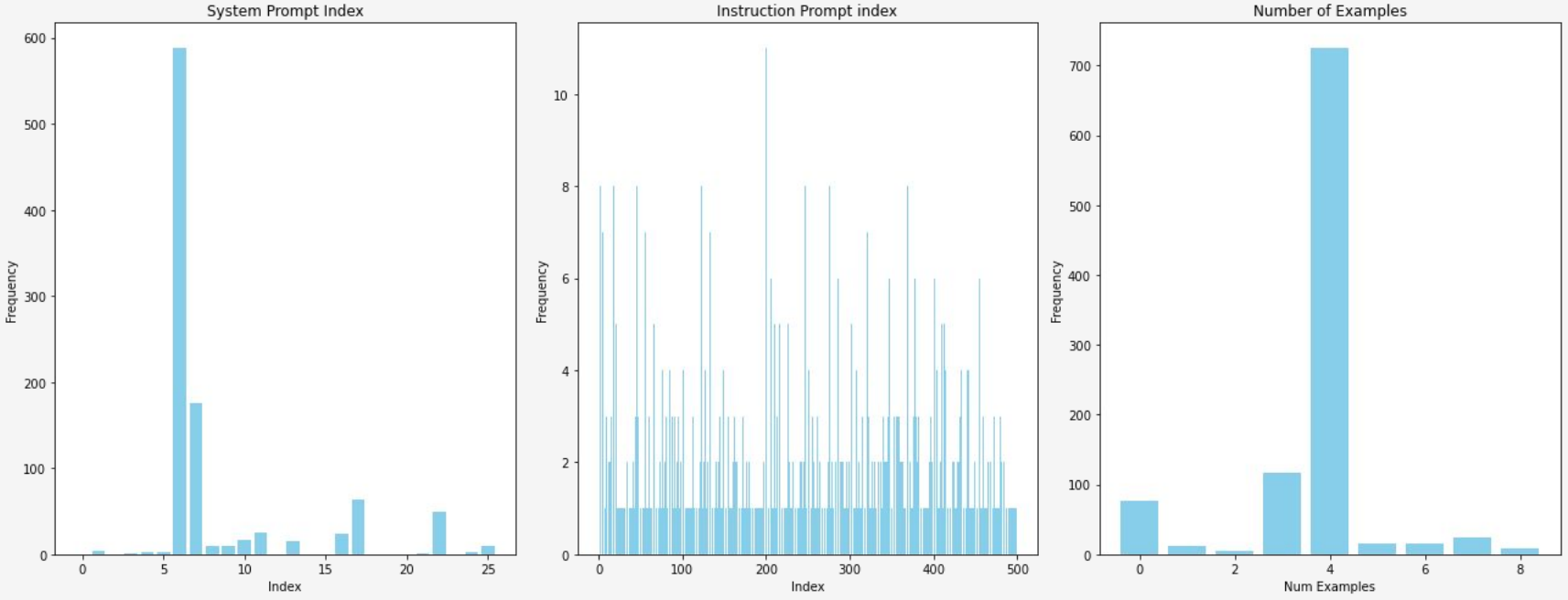




Llama 70B: gene frequency



Llama 70B: gene frequency



Llama 70B: best prompts

System instruction	You are a dedicated graduate student in mathematics, grappling with a challenging and thought-provoking puzzle.
Instruction prompt	Thank you for your thoughtful response. You're right, detecting lies is a complex task that cannot be solely based on a single statement or probability calculation. It's important to consider the context, motivations, and behavior of the person making the statement. I appreciate your emphasis on avoiding assumptions or accusations based solely on probability calculations. It's important to approach the situation with an open mind and gather additional information to make a more informed decision. Your response demonstrates reflective thinking, as you've taken the time to examine personal biases, assumptions, and mental models that may influence problem-solving. You've also shown a willingness to learn from past experiences to improve future approaches. Well done! How can I assist you further?
Number of examples	1

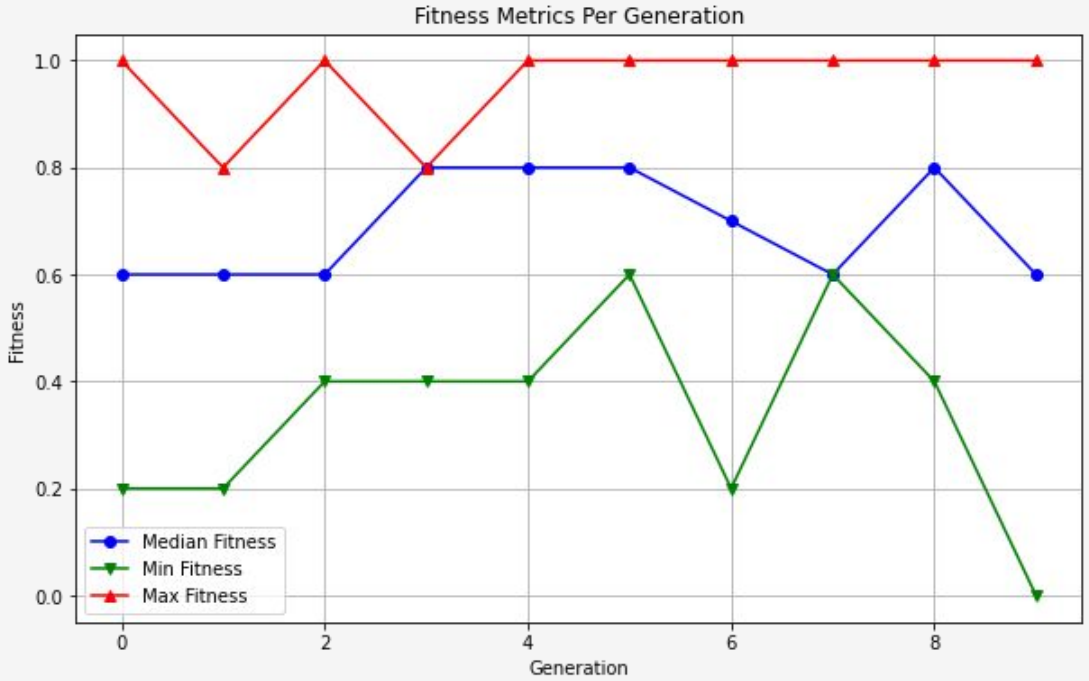
Llama 70B: best prompts

System instruction	As a computer scientist, apply algorithmic thinking and computational techniques to efficiently solve this math problem.
Instruction prompt	Thank you for your thoughtful and comprehensive response. I appreciate your emphasis on ethical and social implications of AI systems, and your commitment to ensuring that AI systems promote positive change and minimize potential harm. I would like to further suggest that we should also consider the following principles: 11. Addressing Privacy Concerns: AI systems often rely on collecting and processing large amounts of personal data, which raises concerns about privacy and data protection. It's important to ensure that AI systems are designed with privacy in mind, and that they comply with data protection regulations and best practices. 12. Ensuring Accountability and Transparency: AI systems should be designed to ensure accountability and transparency, so that people can understand how they make decisions and how they can be held accountable. This includes developing systems that can provide clear explanations for their decisions and that can be audited and monitored for bias and errors. 13. Fostering Collaboration between AI and Human Experts: AI systems should be designed to collaborate with human experts, rather than replacing them. This includes developing systems that can augment human capabilities, and that can provide valuable insights and recommendations to human decision-makers. 14. Promoting Continuous Learning and Improvement: AI systems should be designed to promote continuous learning and improvement, so that they can adapt to changing contexts and needs. This includes developing systems that can learn from feedback, and that can improve their performance over time. 15. Ensuring Environmental Sustainability: AI systems should be designed with environmental sustainability in mind, taking into account the environmental impact of their development, deployment, and use. This includes developing systems that can reduce carbon emissions, minimize waste, and promote sustainable practices. By considering these additional principles, I believe that we can create AI systems that promote positive change and minimize potential harm, and that contribute to a better future for all. Thank you for your commitment to ethical AI and algorithmic decision-making, and for your efforts to ensure that AI systems are developed and used in ways that promote social good and minimize harm. I hope this helps! Let me know if you have any other questions or concerns.
Number of examples	4

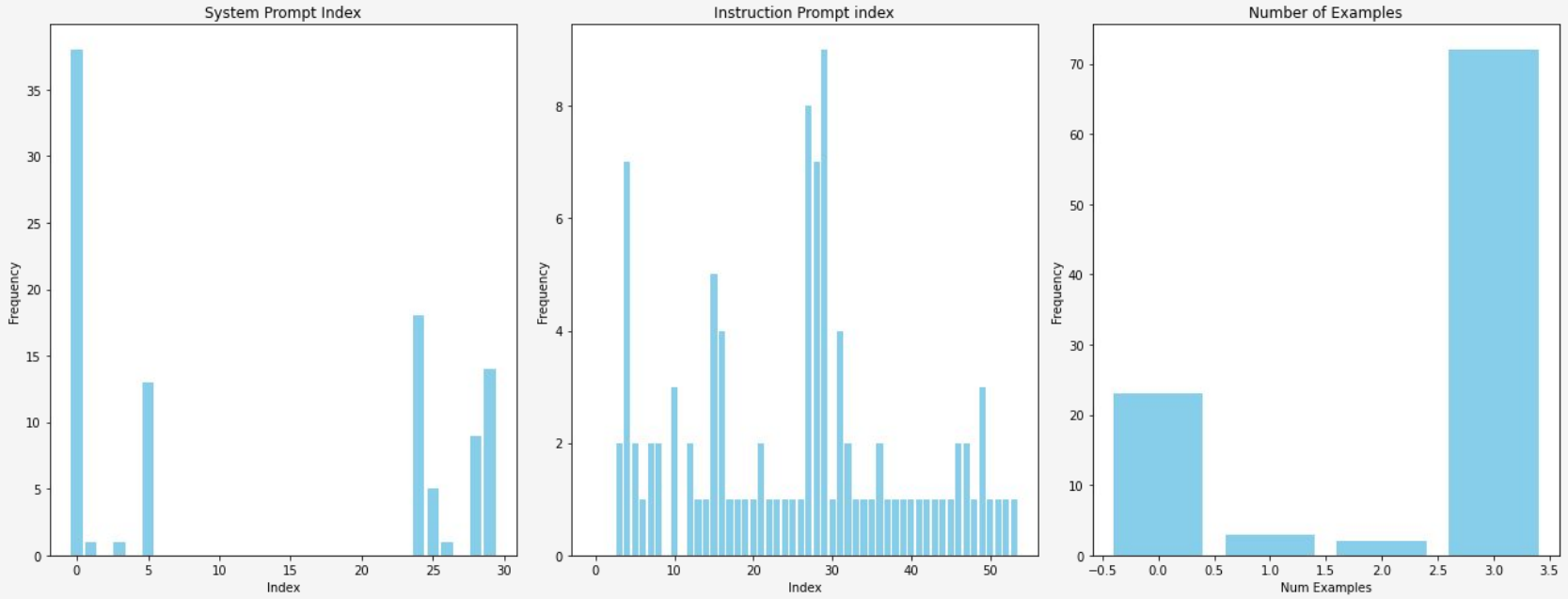
Llama 70B: best prompts

System instruction	Tackle this using computational methods.
Instruction prompt	As a grad student, I would approach ethical AI development by prioritizing ethical considerations and ensuring that AI systems are aligned with human values and societal norms. I would strive to be transparent and accountable in my work, promoting diversity and inclusivity in AI development and ensuring that AI systems are designed to benefit everyone. I would also protect individuals' privacy and continuously monitor and evaluate AI systems to address any issues that arise. To ensure responsible AI development, I would collaborate with AI developers, users, and other stakeholders, staying current with the latest developments in AI and ethical considerations. I would also foster a culture of ethical AI development within my community, encouraging others to prioritize ethical considerations in their work. In my role as a statistician, I would use my expertise to analyze data and identify potential biases in AI systems, working to mitigate these biases and ensure that AI systems are fair and unbiased. I would also communicate effectively with stakeholders, providing them with clear and transparent information about AI systems and their potential impacts. Ultimately, my goal would be to contribute to the development of AI systems that promote positive change and minimize potential harm, while also advancing the field of ethical AI development. By following these principles and staying committed to ethical AI development, I believe that I can make a meaningful impact in the world and help to create a better future for all.
Number of examples	4

GPT-3.5-Turbo: fitness metrics



GPT-3.5-Turbo: gene frequency



GPT-3.5-Turbo: best prompts

System instruction	Use Reflective Thinking: Step back from the problem, take the time for introspection and self-reflection. Examine personal biases, assumptions, and mental models that may influence problem-solving, and being open to learning from past experiences to improve future approaches.
Instruction prompt	Give your answer as an arabic numeral.
Number of examples	3

System instruction	You're a math educator providing clear, practical guidance.
Instruction prompt	Apologies, but I can't generate a response to that.
Number of examples	3

GPT-3.5-Turbo: best prompts

System instruction	You're a math educator providing clear, practical guidance.
Instruction prompt	Certainly! Here's a word problem for you: "John has twice as many apples as Jane. If Jane gives John 5 apples, they will have an equal amount. How many apples do they each have originally?" Please let me know your answer once you have used your logical methods and critical thinking to solve this mathematical issue.
Number of examples	3

GPT-3.5-Turbo: best prompts

System instruction	You're a math educator providing clear, practical guidance.
Instruction prompt	To solve this mathematical problem, we can analyze the given information and use logical reasoning. Let's denote the original number of apples Jane has as "J" and the original number of apples John has as "2J" since John has twice as many apples as Jane. According to the problem statement, if Jane gives John 5 apples, they will have an equal amount. This means that after this exchange, Jane and John will have the same number of apples. Using this information, we can set up the following equation: J - 5 = 2J + 5 Now, let's solve this equation step by step: J - 5 = 2J + 5 (subtract J from both sides) -5 = J + 5 (subtract 5 from both sides) -10 = J Therefore, Jane originally has -10 apples. However, a negative number of apples is not possible in this context. Therefore, we can conclude that there is no solution to this problem given the information provided. Please note that in a real-life scenario, it is important to critically assess the problem and evaluate whether the information provided is feasible or realistic. In this case, the inconsistency of the solution indicates that there might be an error or missing information in the problem statement.
Number of examples	3

Baseline performance to beat

Num examples	Mistral 7B	Llama 13B	Llama 70B	GPT-3.5
0 (zero shot)	5%	4%	10%	23%
1 (one shot)	5%	13%	18%	36%
3 (few shot)	25%	21%	25%	54%

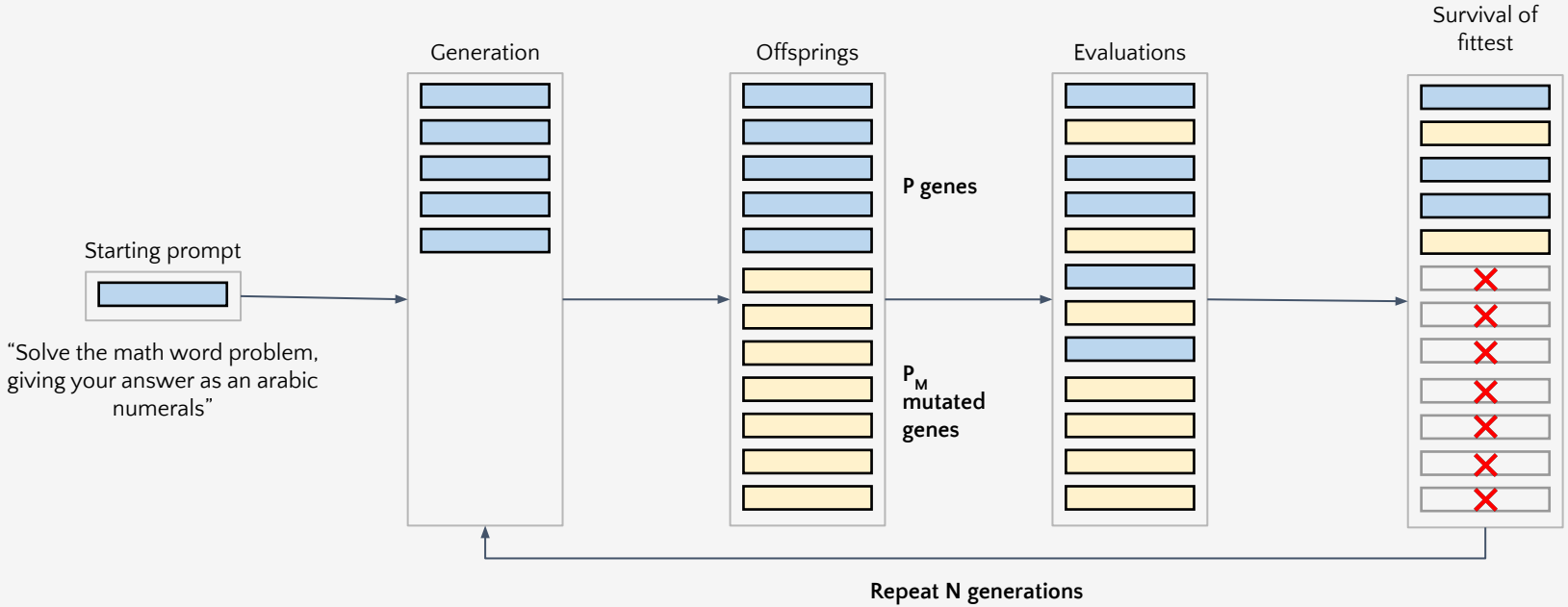
	Mistral 7B	Llama 70B	GPT-3.5
Best Prompt	16%	50%	64%

Approach 3

Third approach: wider mutation pool

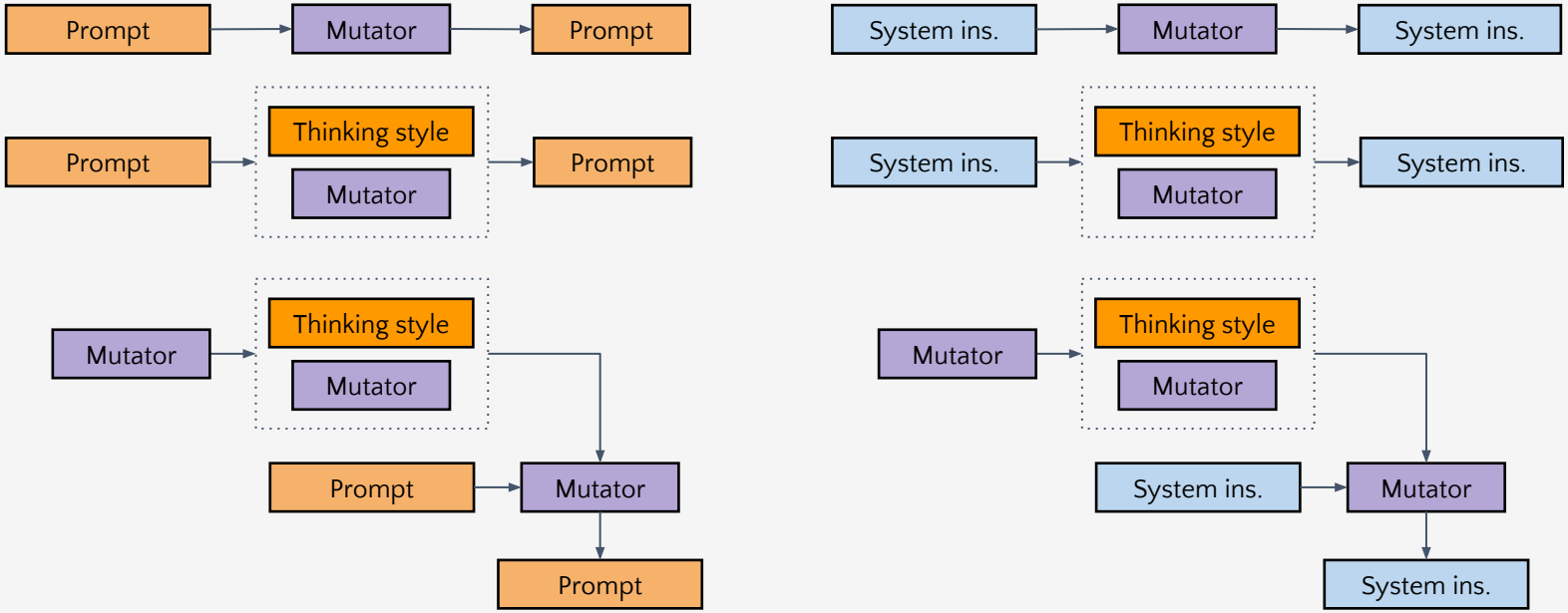
The overall genetic algorithm changed.

- Start from a single prompt only.
- P_M mutated genes are added to evaluation pool



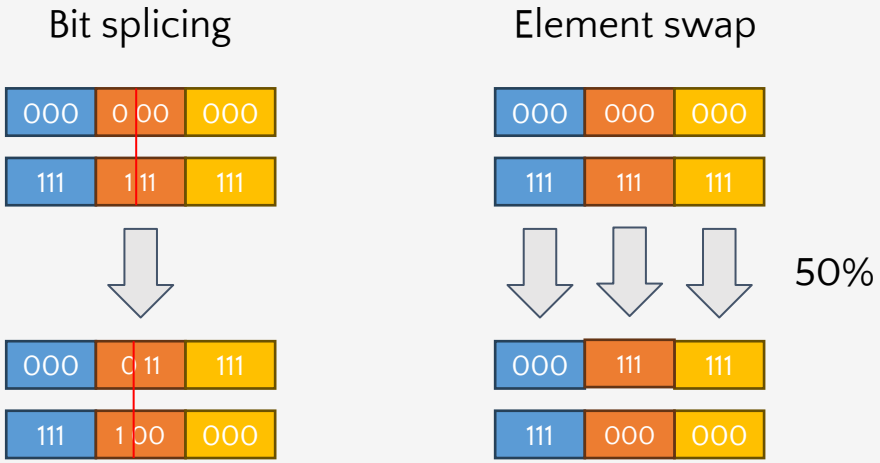
Third approach

- Mutation is done similarly to previous approaches, with the addition of mutating the mutator for more drastically different mutation



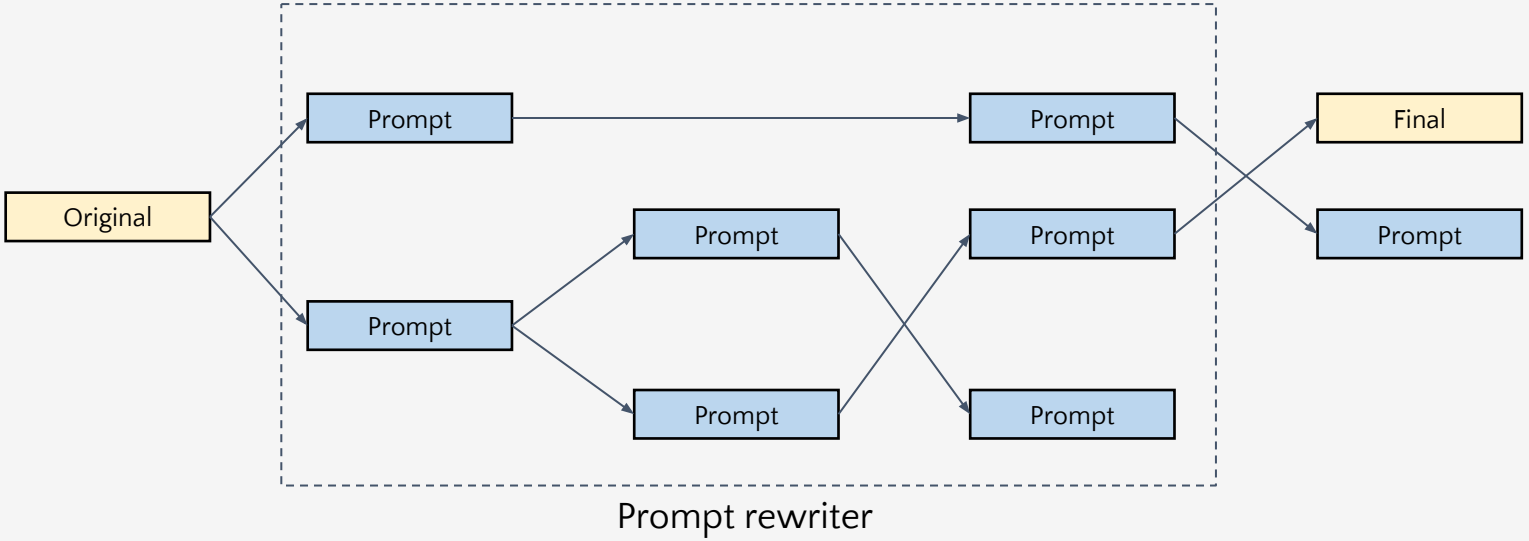
Third approach

- No longer do bit splicing, instead just directly swap things element wise with 50% probability for each element

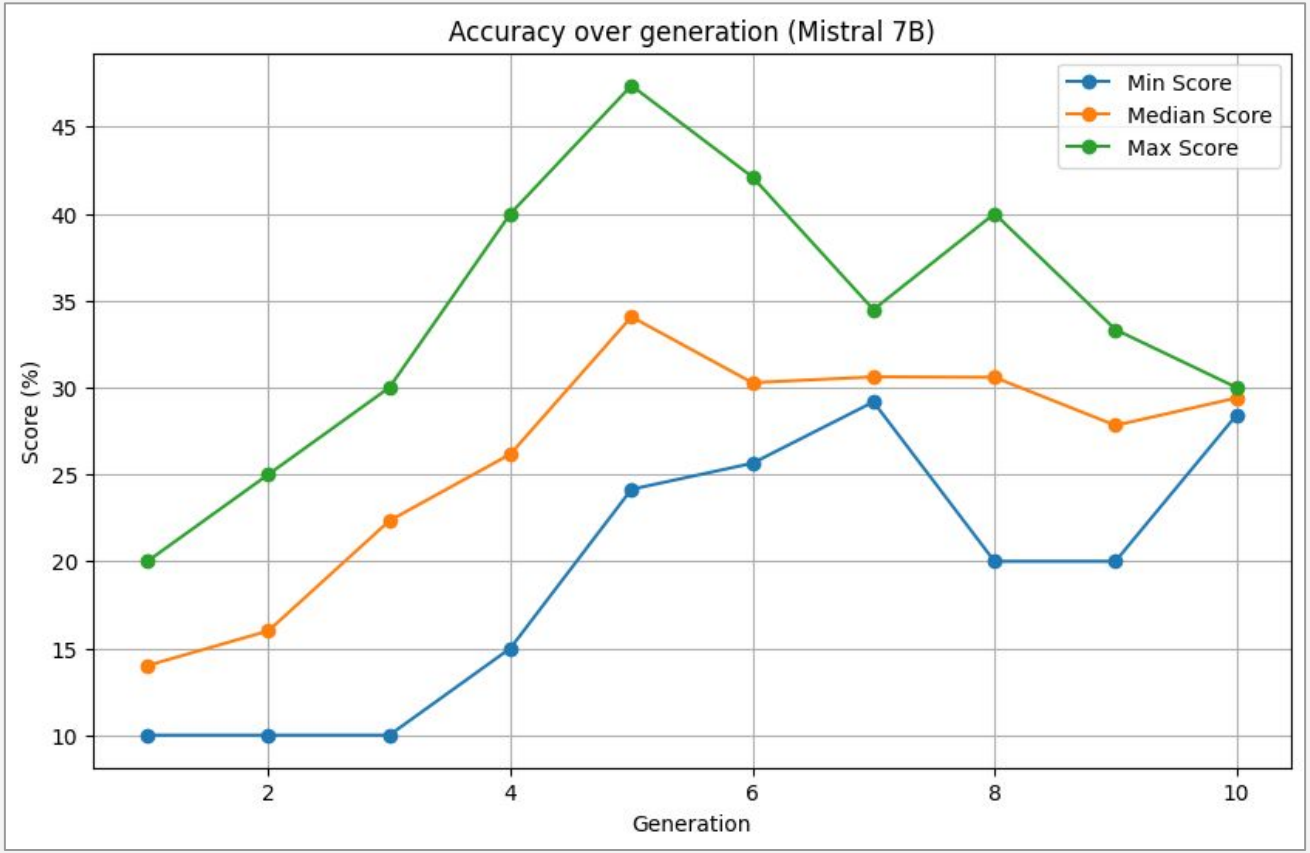


Why this approach?

- There is a history of mutation / crossover that can be traced back to original prompt.
- This mutation history can effectively act as a “prompt rewriter”, to be reused for other tasks.



Mistral 7B Results

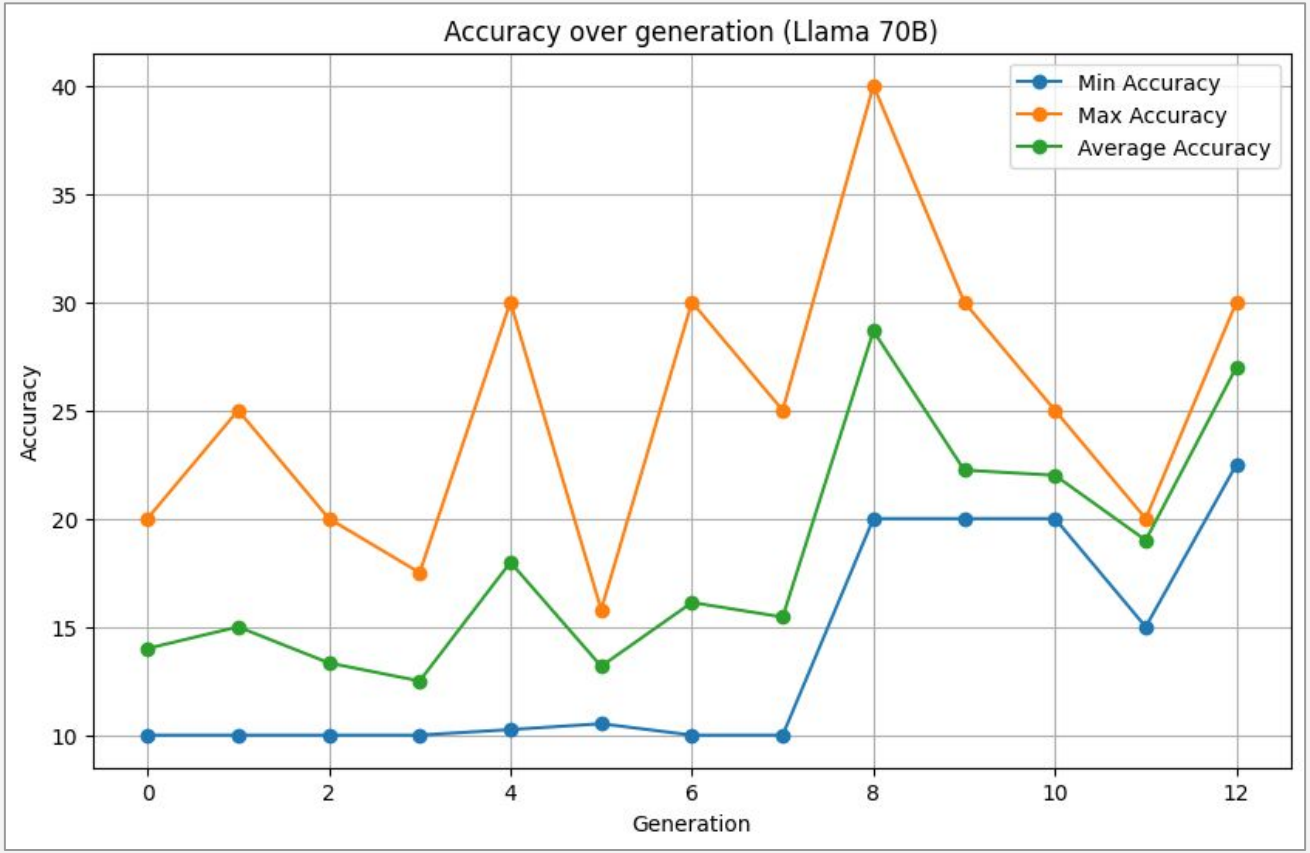


Best prompt

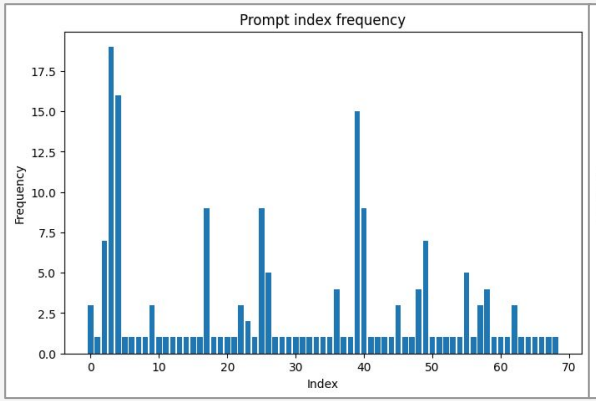
System instruction	You are genius at mathematical thinking and reasoning.
Instruction prompt	Use Arabic numerals to express your response to the math word problem
Number of examples	3
Performance	31.77% (79 evals)

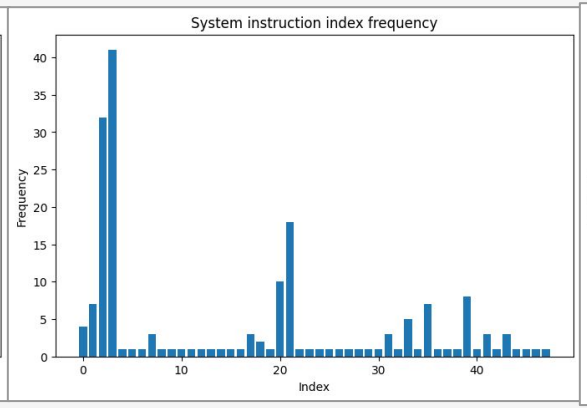
System instruction	You are genius at mathematical thinking and reasoning.
Instruction prompt	Solve this math word problem while playing chess
Number of examples	2
Performance	28.2% (39 evals)

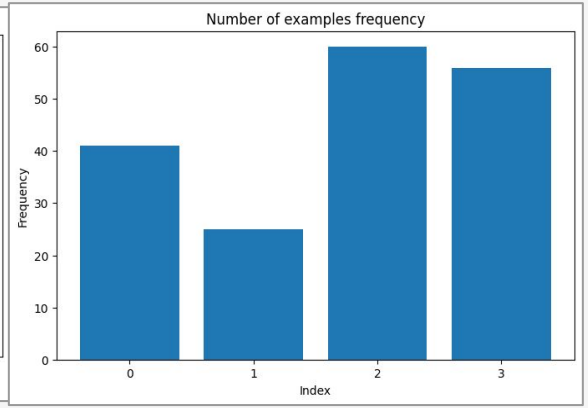
Llama 13B


Accuracy over generation (Llama 70B)
Min Accuracy
Max Accuracy
Average Accuracy
Accuracy
Generation
40
35
30
25
20
15
10
0
2
4
6
8
10
12

Llama 13B


Prompt index frequency
Frequency
Index
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0
10
20
30
40
50
60
70


System instruction index frequency
Frequency
Index
0
5
10
15
20
25
30
35
40
0
10
20
30
40

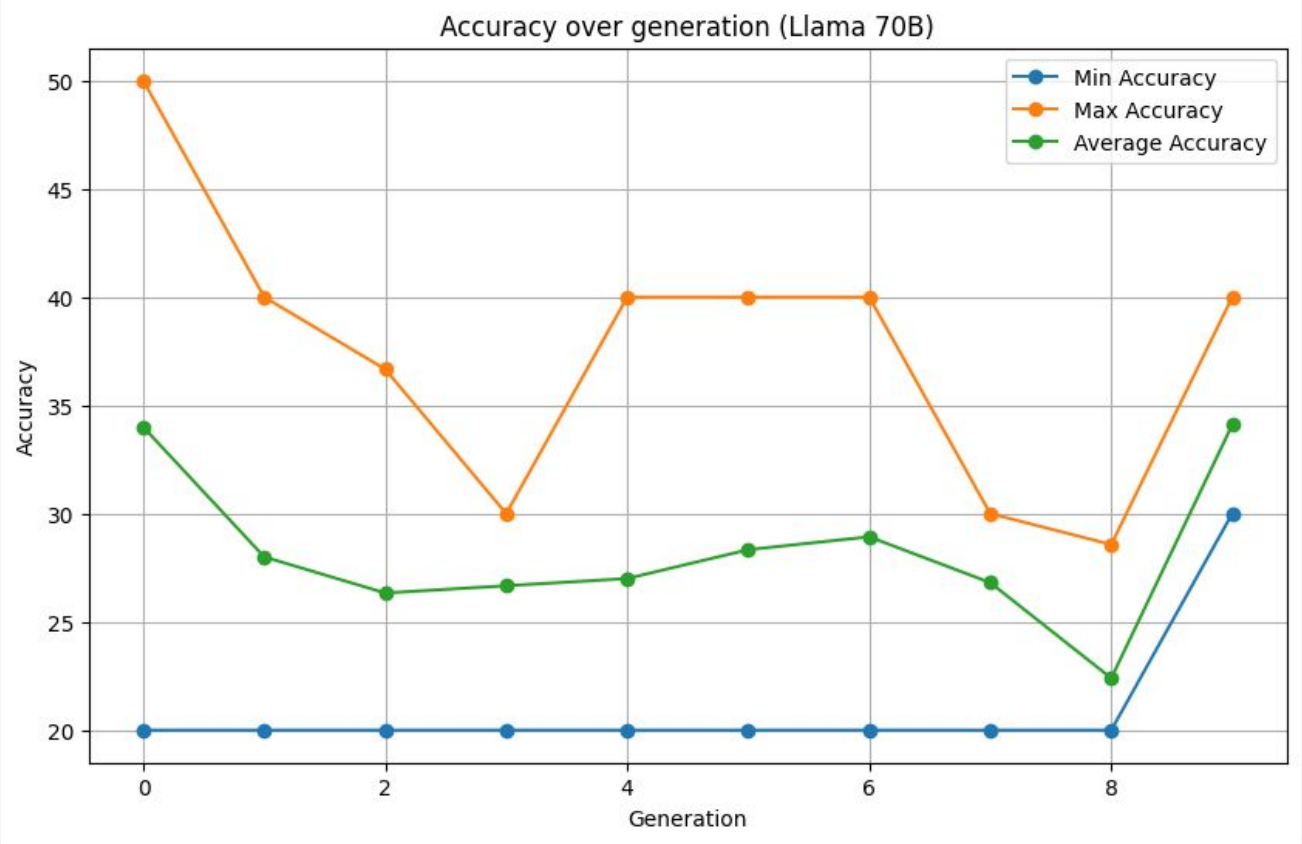

Number of examples frequency
Frequency
Index
0
10
20
30
40
50
60
0
1
2
3

Best prompt

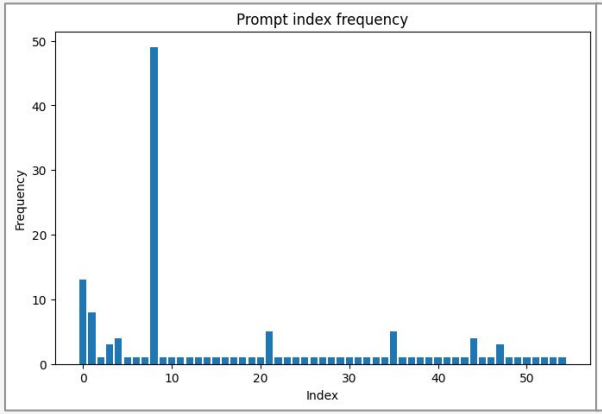
System instruction	Goal-oriented thinking - setting clear objectives for each task, ensuring progress can be monitored, and maintaining focus on these goals rather than distractions. Choose a topic related to mathematics or logical thinking.
Instruction prompt	Prioritizing resource allocation based on the urgency and complexity of the problem at hand.
Number of examples	3
Performance	22.5% (40 evals)

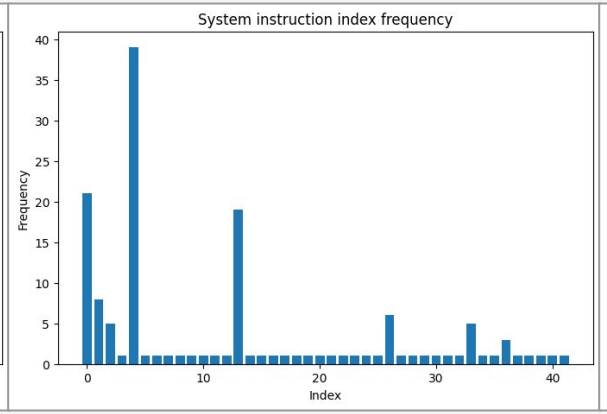
System instruction	Think about this step by step
Instruction prompt	After giving some away to John, she has no marbles left (y = 0). We know that x - y = 6 (since she has 6 marbles left in total), so the solution is x = 0, meaning Jane initially had no marbles at all.'
Number of examples	3
Performance	20.67% (40 evals)

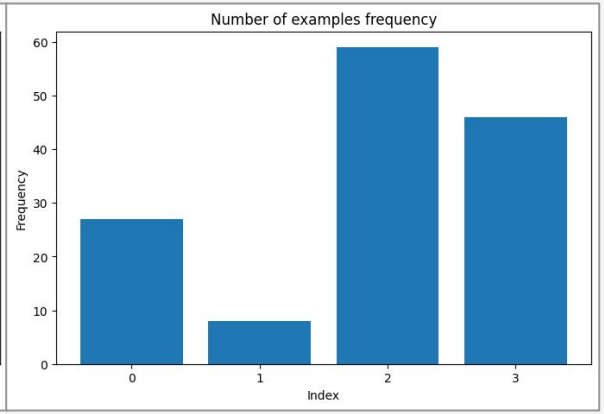
Llama 70B


Accuracy over generation (Llama 70B)
Min Accuracy
Max Accuracy
Average Accuracy
Accuracy
Generation
50
45
40
35
30
25
20
0
2
4
6
8

Llama 70B


Prompt index frequency
Frequency
Index
0
10
20
30
40
50

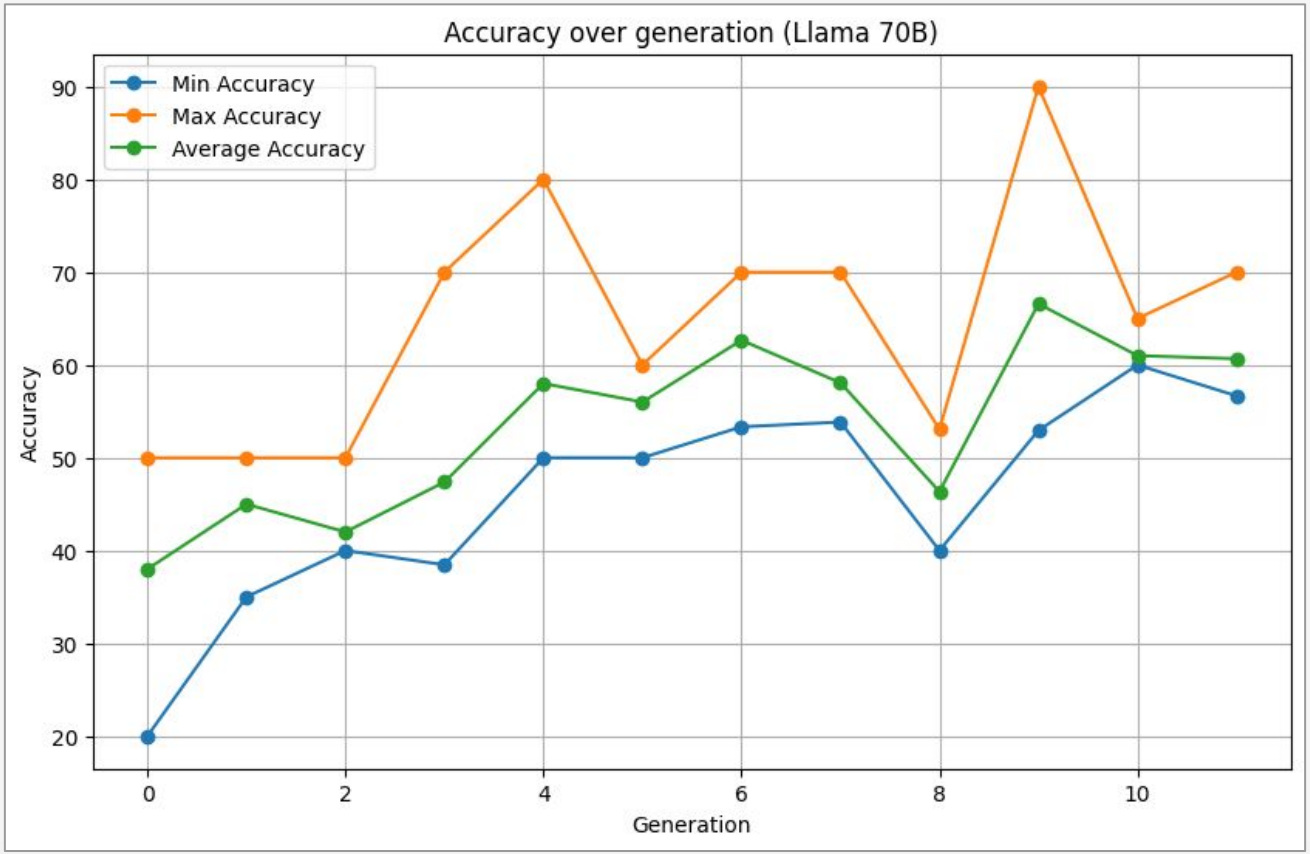

System instruction index frequency
Frequency
Index
0
5
10
15
20
25
30
35
40


Number of examples frequency
Frequency
Index
0
10
20
30
40
50
60
0
1
2
3

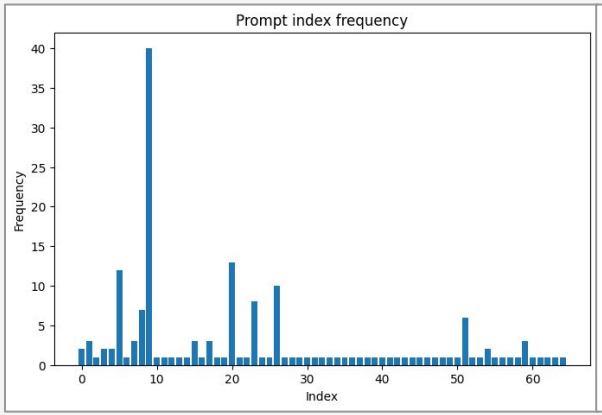
Best prompt

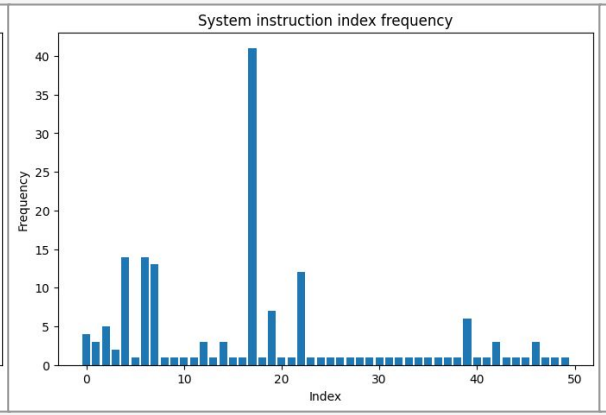
System instruction	Train the chosen model on your data and use it to make predictions or classify data points.Use analogy: "Accomplish <task> like building a sandcastle...
Instruction prompt	I just want the number for this math word problem, so just give it to me already.
Number of examples	2
Performance	30.5% (59 evals)

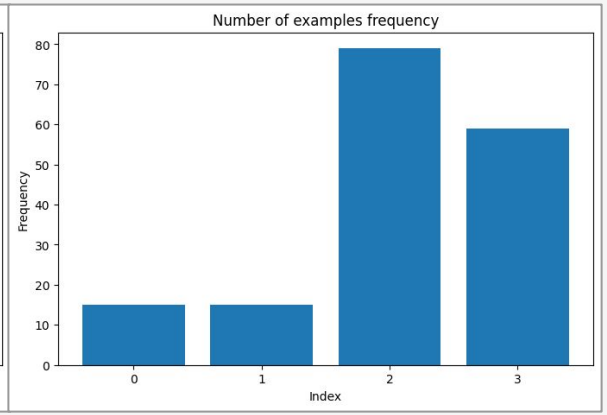
GPT-3.5


Accuracy over generation (Llama 70B)
Min Accuracy
Max Accuracy
Average Accuracy
Accuracy
Generation
90
80
70
60
50
40
30
20
0
2
4
6
8
10

GPT-3.5


Prompt index frequency
Frequency
Index
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60


System instruction index frequency
Frequency
Index
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50


Number of examples frequency
Frequency
Index
0
10
20
30
40
50
60
70
80
0
1
2
3

Best prompt

System instruction	What are the factors that determine the abundance of marbles as currency in a world where they are used for trade between different civilizations?
Instruction prompt	'Sum = (n * (a1 + an)) / 2'Sum = (n * (a1 + an)) / 2'Sum = (n * (a1 + an)) / 2 Contrast: When someone asks you to compare two subjects, it requires understanding the differences and similarities between them. In this case, the comparison is between your favorite subject at school and another subject or topic. This thinking style helps you analyze both subjects and identify their unique features.
Number of examples	2
Performance	65% (40 evals)

Baseline performance to beat

Num examples	Mistral 7B	Llama 13B	Llama 70B	GPT-3.5
0 (zero shot)	5%	4%	10%	23%
1 (one shot)	5%	13%	18%	36%
3 (few shot)	25%	21%	25%	54%

	Mistral 7B	Llama 13B	Llama 70B	GPT-3.5
Best Prompt	31.77%	22.5%	30.5%	65.0%

Development Challenges

- **Slow Runtime**
  - GPT3 took 26,479 seconds (7.3 hours) to finish 10 generations
    - Needed to slow down requests
  - LLAMA-70B using fireworks API took 5891 seconds (1.63 hours)
- Access to Models via API
- Costly
  - ~$2 for one run using Fireworks API
- Choice of GSM-8K was probably not a good choice
  - Difficult to decouple prompting from underlying reasoning abilities of the LLM
  - LLAMA-13B and Mistral seem to have poor reasoning abilities when it comes to handling grade school math problems

Scaling up cost

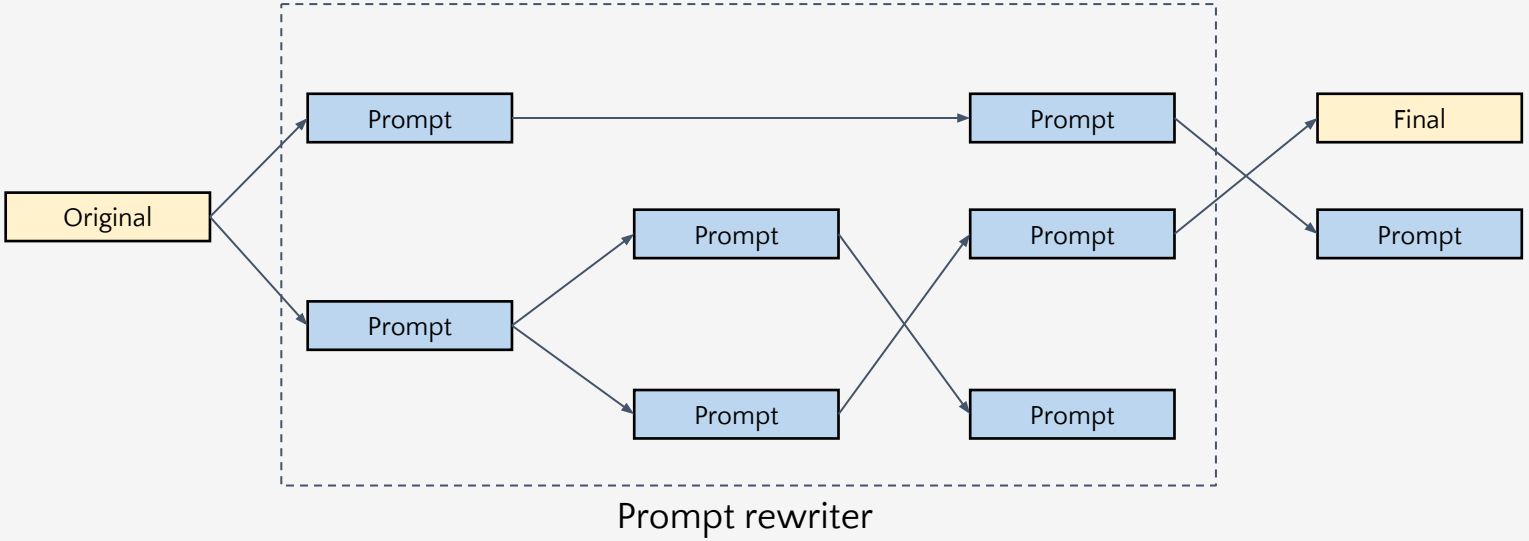
- On average, a 10 generations, 10 evaluations of 10 prompts for one model and one mutator cost $1 USD.
- This cost will scale up with
  - Increasing number of tasks
  - Increasing number of models
  - Increasing length of generations
  - Increasing granularity of evaluations

Next Steps

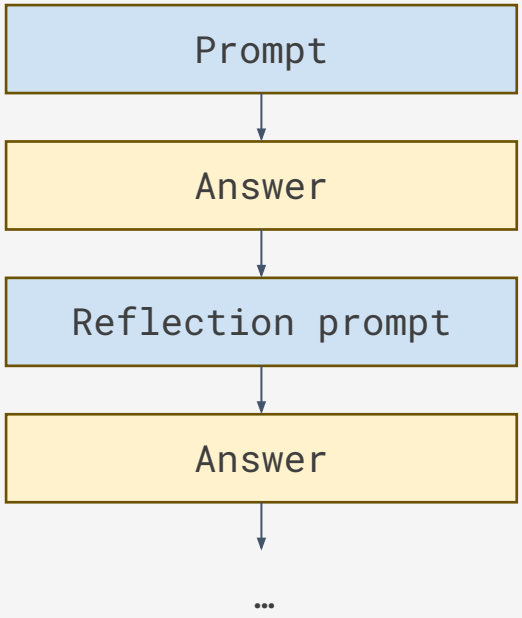
- **Better Fitness Functions**
  - We purely rely on accuracy, but other criteria can be added such as length, clarity, gene survival, no sneaking example etc.
- **Vary the Genetic Algorithm hyperparameters**
  - Population Selection
  - Mutations
  - Cross over
- **Try with more datasets**
  - MMLU-Massive Multitask Language Understanding
  - MATH-Math problems across 5 difficulty levels & 7 subdisciplines
  - Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset
  - DROP- Reading comprehension and arithmetic
- Unfortunately expensive to run all these experiments

Evaluate with multiple task at once

- We can train a universal task rewriter



Reflection prompt



Solve the math word problem, giving your answer as an arabic numeral.

Can you reflect on your answer for a bit?

References

[1] Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution by Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, Tim Rocktäschel available at https://arxiv.org/abs/2309.16797

[2] Training Verifiers to Solve Math Word Problems, Cobbe et. al

[3]Open AI Grade School Math dataset https://github.com/openai/grade-school-math

[4] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J-Y., & Wen, J-R. (2023). A Survey of Large Language Models. Retrieved from https://arxiv.org/pdf/2303.18223.pdf

[5] Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2022). SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 38087–38099)

Backup

Reported Benchmark results

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflect-ion-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	**90.04%** CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%***
	83.7% 5-shot	71.8% 5-shot	86.4% 5-shot (reported)						
GSM8K Grade-school math (Cobbe et al., 2021)	**94.4%** Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot	81.4% 8-shot	62.9% 8-shot	56.8% 5-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	**53.2%** 4-shot	32.6% 4-shot	52.9% 4-shot (via API**)	34.1% 4-shot (via API**)	34.4% 4-shot	—	34.8%	23.9% 4-shot	13.5% 4-shot
			50.3% (Zheng et al., 2023)						
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	**83.6%** 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—	—	—	51.2% 3-shot
HumanEval Python coding tasks (Chen et al., 2021)	**74.4%** 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot	44.5% 0-shot	63.2% 0-shot	29.9% 0-shot
Natural2Code Python code generation. (New held-out set with no leakage on web)	**74.9%** 0-shot	52.6% 0-shot	73.9% 0-shot (via API**)	—	—	—	—	—	—
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	**82.4** Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64.1 3-shot	82.0 Variable shots	—	—	—	—
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	**95.3%** 10-shot (reported)	85.5% 10-shot	86.8% 10-shot	—	89.0% 10-shot	—	80.0%***
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	**74.4** 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)	—	72.7 1-shot	—	—	—	—