

Problem-Solution Overfitting

Observation in AIW: breakdown in easy problem while SOTA in difficult problem -> overfitting in difficult problem? underfitting in easy problem? no generalisation from difficult to easy problem?

Hypothesis: problem-solution overfitting exists

- problem_space: latent space in LLM
- $\Pr(\text{solution}|\text{problem}) = \Pr(\text{problem_space}|\text{problem}) * \Pr(\text{solution}|\text{problem, problem_space})$
- If problem-solution overfitting exists, LLM will look up the problem space, and generate solution conditional on the problem space mostly
- i.e. If $\Pr(\text{solution}|\text{problem}) = \Pr(\text{problem_space}|\text{problem}) * \Pr(\text{solution}|\text{problem_space})$ holds -> $\Pr(\text{solution}|\text{problem_space, problem}) = \Pr(\text{solution}|\text{problem_space})$
- Potential issues arise when **problem provides additional knowledge** about the problem space:
 - LLM gives out answer, even without seeing the question
 - LLM is fairly robust to minor change in details within problem space, but not robust to major change in details outside problem space

Implication:

- If it is problem-solution overfitting contributing to high score in benchmark, benchmark cannot be trusted.
- Breakdown of SOTA for easy question, generalisation is still a fantasy?

GSM8K-NoQ (Test)

- Prompt without Question
 - Standard:
 - Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did he make? Provide then the final answer in following form: "### Answer: . You can answer None if you don't know.
 - No Question
 - Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%.
 - Don't Answer
 - Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. **Don't answer.**

GSM8K-NoQ (Test)

Even without the question, out of all response, 20-30% **last** number can match up to GSM-8K answer in test set. 40-50% of the responses contains the test set answer. In some model like Llama-3.1, phi and Gemma, “answer” appears in the answer, even the prompt has no question at all! Moreover, we see model still gives us answer even we explicitly prompt them don't. Evidence of OverFitting.

model_name	noq_last_match_ans	%	noq_any_match_ans	%	ans_keyword	%	dont_ans_last_match	%	dont_ans_any_match	%	count
claude-3-5-sonnet-20241022	373	28.3	591	44.8	58	4.4	113	8.6	209	15.8	1319
claude-3-haiku-20240307	310	23.5	574	43.5	13	1.0	11	0.8	29	2.2	1319
microsoft/Phi-3-medium-4k-instruct	449	34.0	681	51.6	947	71.8	296	22.4	508	38.5	1319
meta-llama/Meta-Llama-3.1-8B-Instruct	431	32.7	668	50.6	949	71.9	310	23.5	521	39.5	1319
google/gemma-2-9b-it	349	26.5	538	40.8	567	43.0	7	0.5	13	1.0	1319
google/gemma-2-27b-it	367	27.8	552	41.8	495	37.5	3	0.2	4	0.3	1319
Qwen/Qwen2-7B-Instruct	358	27.1	593	45.0	38	2.9	350	26.5	563	42.7	1319
meta-llama/Meta-Llama-3.1-70B-Instruct	479	36.3	662	50.2	986	74.8	191	14.5	283	21.5	1319
meta-llama/Meta-Llama-3.1-405B-Instruct	471	35.7	649	49.2	840	63.7	172	13.0	270	20.5	1319
Qwen/Qwen2-72B-Instruct	439	33.3	646	49.0	151	11.4	19	1.4	30	2.3	1319
gpt-3.5-turbo	301	22.9	503	38.3	5	0.4	117	8.9	190	14.5	1313
gpt-4o	356	27.1	575	43.8	4	0.3	3	0.2	8	0.6	1313
gpt-4o-mini	372	28.3	615	46.8	10	0.8	2	0.2	3	0.2	1313

Failure Case

```
{'dont_answer': 'Marty has 100 centimeters of ribbon that he must cut into 4 '  
    'equal parts. Each of the cut parts must be divided into 5 '  
    'equal parts. Don't answer.',  
'dont_answer_result': "I'll help you solve this step by step, but I won't "  
    'give the final answer.\n'  
    '\n'  
    "Let's break it down:\n"  
    '1. First, Marty needs to cut 100 cm into 4 equal '  
    'parts\n'  
    '    * Each part will be  $100 \div 4$  centimeters\n'  
    '\n'  
    '2. Then, each of these parts needs to be divided into '  
    '5 equal pieces\n'  
    '    * Each of these smaller pieces will be  $(100 \div 4) \div '$   
    '5 centimeters\n'  
    '\n'  
    'You can solve this by:\n'  
    '1. Dividing 100 by 4 to find the length of each first '  
    'cut\n'  
    '2. Then dividing that result by 5 to find the length '  
    'of each final piece',  
'model_name': 'claude-3-5-sonnet-20241022',
```

```
'model_name': 'claude-3-5-sonnet-20241022',  
'no_question': 'Marty has 100 centimeters of ribbon that he must cut into 4 '  
    'equal parts. Each of the cut parts must be divided into 5 '  
    'equal parts.',  
'no_question_result': 'Let me help you solve this step by step.\n'  
    '\n'  
    '1) First, let's find the length of each of the 4 equal "  
    'parts\n'  
    '    * Total length = 100 centimeters\n'  
    '    * Number of parts = 4\n'  
    '    * Length of each part =  $100 \div 4 = 25$  centimeters\n'  
    '\n'  
    '2) Then, each of these 25 centimeter parts must be '  
    'divided into 5 equal parts\n'  
    '    * Length of each part = 25 centimeters\n'  
    '    * Number of smaller parts = 5\n'  
    '    * Length of each smaller part =  $25 \div 5 = 5$  '  
    'centimeters\n'  
    '\n'  
    'Therefore, after all the cutting is done:\n'  
    '* Each small piece will be 5 centimeters long\n'  
    '* There will be 5 small pieces from each of the 4 '  
    'larger pieces\n'  
    '* Total number of small pieces =  $4 \times 5 = 20$  pieces\n'  
    '\n'  
    'The answer is 5 centimeters for each final piece.'}
```

GSM8K-Who (Test)

- Change the question noun to add new context:
 - Standard:
 - Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did **he** make? Provide then the final answer in following form: "### Answer: . You can answer None if you don't know.
 - Sibling
 - Josh decides to try flipping a house. He buys a house or \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did **Josh's** brother make? Provide then the final answer in following form: "### Answer: ". You can answer None if you don't know.
 - Random person
 - Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did **Judy** make? Provide then the final answer in following form: "### Answer: ". You can answer None if you don't know.
- Instruction of "You can answer None if you don't know." is added to guide model to abstain from answering.

GSM8K-Who (Test)

On average, only 10-20% outputs None despite higher percentage of mention of random person. For sibling perturbation, result is higher, but still lower than the mention. The fact that frequency of None < frequency of Mention is alarming because it implies the model is aware of new entity but still not able to give None.

model_name	matched_ans	random_matched_ans	%	random_none	%	random_mention	%	sibling_matched_ans	%	sibling_none	%	sibling mention	%	count
gpt-3.5-turbo	297	264	88.9	65	10.5	198	32.1	270	90.9	97	15.7	306	49.6	617
gpt-4o	535	474	88.6	119	19.3	224	36.3	409	76.4	241	39.1	456	73.9	617
claude-3-haiku-20240307	478	432	90.4	71	11.5	192	31.1	407	85.1	103	16.7	437	70.8	617
gpt-4o-mini	538	481	89.4	95	15.4	220	35.7	460	85.5	122	19.8	466	75.5	617
google/gemma-2-27b-it	487	462	94.9	56	9.1	99	16.0	430	88.3	122	19.8	202	32.7	617
claude-3-5-sonnet-20241022	550	458	83.3	192	31.1	245	39.7	367	66.7	350	56.7	412	66.8	617
google/gemma-2-9b-it	511	453	88.6	70	11.3	115	18.6	430	84.1	150	24.3	221	35.8	617
Qwen/Qwen2-7B-Instruct	474	426	89.9	74	12.0	165	26.7	429	90.5	65	10.5	412	66.8	617
meta-llama/Meta-Llama-3.1-8B-Instruct	511	426	83.4	83	13.5	239	38.7	342	66.9	169	27.4	499	80.9	617
microsoft/Phi-3-medium-4k-instruct	519	428	82.5	81	13.1	236	38.2	351	67.6	164	26.6	498	80.7	617
meta-llama/Meta-Llama-3.1-70B-Instruct	557	445	79.9	196	31.8	275	44.6	340	61.0	359	58.2	487	78.9	617
Qwen/Qwen2-72B-Instruct	543	482	88.8	102	16.5	177	28.7	409	75.3	198	32.1	451	73.1	617
meta-llama/Meta-Llama-3.1-405B-Instruct	559	499	89.3	221	35.8	301	48.8	436	78.0	354	57.4	539	87.4	617

Failure Case

```
'model_name': 'gpt-4o',
'random_question': 'Davos bought 2 shirts from the mall that cost $30 each. '
                    'If he bought them with a 40% discount, how much did Alex '
                    'pay for the 2 shirts? Provide then the final answer in '
                    'following form: "### Answer: ". You can answer None if '
                    '"you don't know."',
'random_result': 'First, calculate the total cost of the 2 shirts without the '
                 'discount:\n'
                 '\n'
                 '\\[ 2 \\times 30 = 60 \\]\n'
                 '\n'
                 'Next, calculate the discount amount:\n'
                 '\n'
                 '\\[ 60 \\times 0.40 = 24 \\]\n'
                 '\n'
                 'Subtract the discount from the total cost to find the '
                 'amount Davos paid:\n'
                 '\n'
                 '\\[ 60 - 24 = 36 \\]\n'
                 '\n'
                 '### Answer: 36',
```

```
'sibling_question': 'Davos bought 2 shirts from the mall that cost $30 each. '
                    'If he bought them with a 40% discount, how much did '
                    'Davos's sister pay for the 2 shirts? Provide then the '
                    'final answer in following form: "### Answer: ". You can '
                    '"answer None if you don't know."',
'sibling_result': "To find out how much Davos's sister paid for the 2 shirts, "
                  'we first need to calculate the total cost of the shirts '
                  'before the discount and then apply the discount.\n'
                  '\n'
                  '1. Calculate the total cost of the 2 shirts without the '
                  'discount:\n'
                  '    \\[\n'
                  '        2 \\text{ shirts} \\times \\$30 \\text{ per shirt} = '
                  '\\\$60\n'
                  '    \\]\n'
                  '\n'
                  '2. Calculate the discount amount:\n'
                  '    \\[\n'
                  '        40\\% \\text{ of } \\$60 = 0.40 \\times \\$60 = \\$24\n'
                  '    \\]\n'
                  '\n'
                  '3. Subtract the discount from the total cost to find the '
                  'final amount paid:\n'
                  '    \\[\n'
                  '        \\$60 - \\$24 = \\$36\n'
```

Next Step

- Fluctuation on same prompt (temp=0 now, but still its stochastic)
- More prompt variation
- More ablation study
 - Length of prompt -> more token -> more overfitting
- Formal definition and measure of Problem-Solution Overfitting
- Formal measure of generalisation gap
- Paper :)