1. Task所属领域

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data | Program | OS | IoT | Software | … |
| Ours |  |  |  |  |  |
| Public |  |  |  |  |  |

1. Snapshots风险类型

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data |  | Privacy Leakage | Computer Security | Financial Loss | Property Damage | Physical Health | Data Loss | Illegal Activities | Ethics & Morality | Bias & Offensiveness | Miscellaneous |
| Ours | Train |  |  |  |  |  |  |  |  |  |  |
|  | Test |  |  |  |  |  |  |  |  |  |  |
| Public | Train |  |  |  |  |  |  |  |  |  |  |
|  | Test |  |  |  |  |  |  |  |  |  |  |

1. Snapshots风险来源

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data |  | User Instruction |  |  | Env |  |
|  |  | Normal | Underspecification | Adversarial | Normal | Adversarial |
| Ours | Train |  |  |  | - |  |
|  | Test |  |  |  |  |  |
| Public | Train |  |  |  |  |  |
|  | Test |  |  |  |  |  |

|  |  |  |
| --- | --- | --- |
|  | Normal | Adversarial |
| Normal |  |  |
| Underspecification |  |  |
| Adversarial |  |  |

1. Public

Step1. Cases √

Step2: snapshot √

Step3: sample action

GPT-4 √

GPT-4o √

Claude-3.5 √

Llama3.1-8b √

Llama3.1-70b √

Qwen- -

Step4: evaluation

GPT-4 √

GPT-4o √

Cluade-3.5 √

Llama3.1-8b √

Llama3.1-70b √

Qwen- -

Table1. Safety evaluation rate. (%)

|  |  |  |  |
| --- | --- | --- | --- |
| Model | @1 | @3 | @5 |
| GPT-4 | 28.57% | 14.29% | 8.16% |
| GPT-4o | 34.69% | 30.61% | 26.53% |
| Cluade-3.5 | 30.0% | 26.0% | 26.0% |
| Llama3.1-8b | 20.0% | 14.0% | 12.0% |
| Llama3.1-70b | 20.0% | 18.0% | 18.0% |
| Qwen- |  |  |  |

sample action

GPT-4 √

GPT-4o √

Claude-3.5 √

Llama3.1-8b √

Llama3.1-70b √

Qwen- -

Evaluation

GPT-4 √

GPT-4o √

Cluade-3.5 √

Llama3.1-8b √

Llama3.1-70b √

Qwen- -

Table2. Naïve method safety evaluation rate. (%)

|  |  |  |  |
| --- | --- | --- | --- |
| Model | @1 | @3 | @5 |
| GPT-4 | 20.83% | 12.5% | 10.42% |
| GPT-4o | 38.78% | 28.57% | 24.49% |
| Cluade-3.5 | 35.42% | 31.25% | 29.17% |
| Llama3.1-8b |  |  |  |
| Llama3.1-70b |  |  |  |
| Qwen- |  |  |  |

Table3. Ours method safety evaluation rate (%)

|  |  |  |  |
| --- | --- | --- | --- |
| Model | @1 | @3 | @5 |
| **Self** |  |  |  |
| Llama3.1-8b | 53.19% | 38.20% | 31.91% |
| Llama3.1-70b | 44.0% (48.0%) | 40.0% (42.0%) | 34.0% (42.0%) |
| Qwen- |  |  |  |
| **Cross** |  |  |  |
| Llama3.1-8b |  |  |  |
| Llama3.1-70b |  |  |  |
| Qwen- |  |  |  |

1. Ours

Step1. Cases √

Step2: snapshot √

Step3: sample action

GPT-4 -

GPT-4o -

Claude-3.5 -

Llama3.1-8b √

Llama3.1-70b √

Qwen- -

Step4: evaluation

GPT-4 -

GPT-4o -

Cluade-3.5 -

Llama3.1-8b √

Llama3.1-70b √

Qwen- -

Table4. Safety evaluation rate. (%)

|  |  |  |  |
| --- | --- | --- | --- |
| Model | @1 | @3 | @5 |
| GPT-4 |  |  |  |
| GPT-4o | 17.78% | 8.89% | 6.67% |
| Cluade-3.5 | 54.0% | 54.0% | 54.0% |
| Llama3.1-8b | 12.0% | 6.0% | 4.0% |
| Llama3.1-70b |  |  |  |
| Qwen- |  |  |  |

sample action

GPT-4 -

GPT-4o -

Claude-3.5 -

Llama3.1-8b -

Llama3.1-70b -

Qwen- -

Evaluation

GPT-4 -

GPT-4o -

Cluade-3.5 -

Llama3.1-8b -

Llama3.1-70b -

Qwen- -

Table5. Naïve method safety evaluation rate. (%)

|  |  |  |  |
| --- | --- | --- | --- |
| Model | @1 | @3 | @5 |
| GPT-4 |  |  |  |
| GPT-4o |  |  |  |
| Cluade-3.5 |  |  |  |
| Llama3.1-8b (24/50) | 45.83% | 25.0% | 20.83% |
| Llama3.1-70b (48/50) | 35.42% | 18.75% | 14.58% |
| Qwen- |  |  |  |

Table3. Ours method safety evaluation rate (%)

|  |  |  |  |
| --- | --- | --- | --- |
| Model | @1 | @3 | @5 |
| Llama3.1-8b |  |  |  |
| Llama3.1-70b | 40.81% | 30.61% | 28.57% |
| Qwen- |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Method | Ours | | | Public | | |
| @1 | @3 | @5 | @1 | @3 | @5 |
| GPT-4 | Baseline | 17.39% | 13.04% | 13.04% | 28.57% | 14.29% | 8.16% |
| GPT-4o | Baseline | 17.78% | 8.89% | 6.67% | 34.69% | 30.61% | 26.53% |
| Cluade-3.5 | Baseline | 54.0% | 54.0% | 54.0% | 30.0% | 26.0% | 26.0% |
| Llama3.1-8b | Baseline | 12.0% | 6.0% | 4.0% | 20.0% | 14.0% | 12.0% |
| Llama3.1-70b | Baseline | 16.32% | 12.24% | 10.20% | 20.0% | 18.0% | 18.0% |
| GPT-4 | Naïve | 30.6% | 24.49% | 24.49% | 20.83% | 12.5% | 10.42% |
| GPT-4o | Naïve | 36.73% | 36.73% | 32.65% | 38.78% | 28.57% | 24.49% |
| Cluade-3.5 | Naïve | 52.0% | 52.0% | 32.0% | 35.42% | 31.25% | 29.17% |
| Llama3.1-8b | Naïve | 28.0% | 18.0% | 26.0% | 28.0% | 12.0% | 10.0% |
| Llama3.1-70b | Naïve | 22.45% | 10.20% | 8.16% | 26.0% | 20.0% | 18.0% |
| Llama3.1-8b | Ours (self) | 53.06% | 36.73% | 32.65% | 56.25% | 39.58% | 35.42% |
| Llama3.1-70b | Ours (self) | 44.0% | 40.0% | 38.0% | 44.0% | 42.0% | 42.0% |
| Llama3.1-8b | Ours (Cross) | 60.0% | 58.0% | 56.0% | 69.39% | 67.35% | 65.31% |
| Llama3.1-70b | Ours (Cross) |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | model | baseline |  |  | Naive |  |  | Ours |  |  |
|  |  | @1 | @2 | @3 | @1 | @2 | @3 | @1 | @2 | @3 |
| Ours | GPT-4 | 13.04% | 13.04% | 13.04% | 30.6% | 24.49% | 24.49% | - | - | - |
|  | GPT-4o | 17.78% | 8.89% | 6.67% | 36.73% | 36.73% | 32.65% | - | - | - |
|  | Cluade-3.5 | 26.67% | 20.0% | 13.0% | 37.78% | 26.67% | 26.67% | - | - | - |
|  | Gemini-1.5-pro | 28.0% | 26.0% | 24.0% | 42.0% | 36.0% | 34.0% |  |  |  |
|  | Llama3.1-8b | 12.0% | 6.0% | 4.0% | 28.0% | 18.0% | 26.0% | 62.0% | 60.0% | 56.0% |
|  | Llama3.1-70b | 16.32% | 12.24% | 10.20% | 22.45% | 10.20% | 8.16% | 64.0% | 60.0% | 58.0% |
|  | Qwen2.5-7b | 20.0% | 14.0% | 12.0% | 18.0% | 16.0% | 12.0% | 68.0% | 64.0% | 62.0% |
|  | Qwn2.5-14b | 18.0% | 16.0% | 14.0% | 28.0% | 26.0% | 22.0% |  |  |  |
|  | Glm-9b | 26.53% | 20.41% | 12.24% | 24.0% | 22.0% | 20.0% | 76.0% | 72.0% | 70.0% |
| Public | GPT-4 | 28.57% | 14.29% | 8.16% | 20.83% | 12.5% | 10.42% | - | - | - |
|  | GPT-4o | 34.69% | 30.61% | 26.53% | 38.78% | 28.57% | 24.49% | - | - | - |
|  | Cluade-3.5 | 30.0% | 26.0% | 26.0% | 35.42% | 31.25% | 29.17% | - | - | - |
|  | Gemini-1.5-pro | 38.0% | 34.0% | 32.0% | 38.0% | 36.0% | 34.0% |  |  |  |
|  | Llama3.1-8b | 20.0% | 14.0% | 12.0% | 28.0% | 12.0% | 10.0% | 58.0% | 56.0% | 54.0% |
|  | Llama3.1-70b | 20.0% | 18.0% | 18.0% | 26.0% | 20.0% | 18.0% | 64.0% | 64.0% | 58.0% |
|  | Qwen2.5-7b | 32.0% | 26.0% | 24.0% | 30.0% | 26.0% | 22.0% | 74.0% | 72.0% | 68.0% |
|  | Qwn2.5-14b | 28.0% | 26.0% | 22.0% | 24.0% | 20.0% | 18.0% |  |  |  |
|  | Glm-9b | 36.0% | 34.0% | 30.0% | 38.0% | 34.0% | 32.0% | 78.0% | 76.0% | 76.0% |

1. Ours vs. Reflection

Table 6. Reflection on Ours

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model |  | 1 | 2 | 3 |  |  |
| Llama3.1-8b | R@1 | 44.89% | 48.98% | 36.73% |  |  |
|  | R@2 | 59.18% | 51.02% | 57.14% |  |  |
|  | R@3 | 63.26% | 61.22% | 68.09% |  |  |
|  | R@4 | 72.92% | 61.22% | 69.39% |  |  |
|  | R@5 | 75.51% | 75.51% | 73.47% |  |  |
|  | R@0 | 12.0% | 6.0% | 4.0% |  |  |
|  | Trained | 62.0% | 60.0% | 56.0% |  |  |
| Llama3.1-70b | R@1 | 74.0% | 72.0% | 74.0% |  |  |
|  | R@2 | 76.0% | 78.0% | 78.0% |  |  |
|  | R@3 | 80.0% | 84.0% | 78.0% |  |  |
|  | R@4 | 82.0% | 84.0% | 80.0% |  |  |
|  | R@5 | 86.0% | 86.0% | 82.0% |  |  |
|  | R@0 | 16.32% | 12.24% | 10.20% |  |  |
|  | Trained | 64.0% | 60.0% | 58.0% |  |  |
| Qwen2.5-7b | R@1 | 63.3% | 67.3% | 67.3% |  |  |
|  | R@2 | 73.5% | 77.6% | 81.6% |  |  |
|  | R@3 | 77.6% | 69.4% | 75.5% |  |  |
|  | R@4 | 81.6% | 85.7% | 81.6% |  |  |
|  | R@5 | 83.7% | 79.6% | 83.3% |  |  |
|  | R@0 | 20.0% | 14.0% | 12.0% |  |  |
|  | Trained | 68.0% | 64.0% | 62.0% |  |  |
| Glm4-9b | R@1 | 68.0% | 76.0% | 74.0% |  |  |
|  | R@2 | 80.0% | 84.0% | 84.0% |  |  |
|  | R@3 | 86.0% | 88.0% | 88.0% |  |  |
|  | R@4 | 84.0% | 90.0% | 88.0% |  |  |
|  | R@5 | 88.0% | 86.0% | 86.0% |  |  |
|  | R@0 | 26.53% | 20.41% | 12.24% |  |  |
|  | Trained | 76.0% | 72.0% | 70.0% |  |  |
|  |  |  |  |  |  |  |

Table 7.Reflection on Public

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model |  | 1 | 2 | 3 |  |
| Llama3.1-8b | R@1 | 36.0% | 48.0% | 44.0% |  |
|  | R@2 | 42.0% | 56.0% | 62.0% |  |
|  | R@3 | 44.0% | 64.0% | 62.0% |  |
|  | R@4 | 48.0% | 68.0% | 70.0% |  |
|  | R@5 | 50.0% | 70.0% | 74.0% |  |
|  | R@0 | 20.0% | 14.0% | 12.0% |  |
|  | Trained | 58.0% | 56.0% | 54.0% |  |
| Llama3.1-70b | R@1 | 70.0% | 68.0% | 66.0% |  |
|  | R@2 | 78.0% | 70.0% | 72.0% |  |
|  | R@3 | 82.0% | 78.0% | 80.0% |  |
|  | R@4 | 88.0% | 80.0% | 86.0% |  |
|  | R@5 | 90.0% | 82.0% | 90.0% |  |
|  | R@0 | 20.0% | 18.0% | 18.0% |  |
|  | Trained | 64.0% | 64.0% | 58.0% |  |
| Qwen2.5-7b | R@1 | 62.0% | 58.0% | 63.0% |  |
|  | R@2 | 66.0% | 59.2% | 63.4% |  |
|  | R@3 | 72.0% | 74.0% | 75.5% |  |
|  | R@4 | 76.0% | 75.5% | 69.4% |  |
|  | R@5 | 66.0% | 74.0% | 72.0% |  |
|  | R@0 | 32.0% | 26.0% | 24.0% |  |
|  | Trained | 74.0% | 72.0% | 68.0% |  |
| Glm4-9b | R@1 | 76.0% | 78.0% | 70.0% |  |
|  | R@2 | 72.0% | 88.0% | 84.0% |  |
|  | R@3 | 90.0% | 92.0% | 86.0% |  |
|  | R@4 | 96.0% | 90.0% | 90.0% |  |
|  | R@5 | 88.0% | 94.0% | 92.0% |  |
|  | R@0 | 36.0% | 34.0% | 30.0% |  |
|  | Trained | 78.0% | 76.0% | 76.0% |  |
|  |  |  |  |  |  |

1. Evaluation

Table 7. Human vs. Ours

|  |  |  |
| --- | --- | --- |
|  | Ours Safety | Ours Unsafety |
| Human Safety | 9 (45%) | 0 (0%) |
| Human Unsafety | 1 (5%) | 10 (50%) |

Table 8. Human vs. GPT-4

|  |  |  |
| --- | --- | --- |
|  | GPT-4 Safety | GPT-4 Safety |
| Human Safety | 9 (45%) | 0 (0%) |
| Human Unsafety | 6 (30%) | 5 (25%) |

Table 9. Human vs. Multi-LLM Vote

|  |  |  |
| --- | --- | --- |
|  | Multi-LLM Vote Safety | Multi-LLM Vote Safety |
| Human Safety | 9 | 0 |
| Human Unsafety | 6 | 5 |

Table 10. Ours

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Num\ Sample | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 |  |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |  |  |  |  |

Table 11. GPT-4

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Num\ Sample | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 |  |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |  |  |  |  |

Table 12. Multi-Model Vote

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Num\ Sample | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 |  |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |  |  |  |  |

1. Useful

|  |  |  |
| --- | --- | --- |
|  | Ours | Public |
| Llama-8b | 30.0% | 40.0% |
| Llama-70b | 87.76% | 78.0% |
| Llama-8b-trained | 74.0% | 91.84% |
| Llama-70b-trained | 100.0% | 98.0% |
| Qwen-7b | 38.0% | 44.0% |
| Glm-9b | 40.0% | 52.0% |
| Qwen-7b-trained | 72.0% | 76.0% |
| Glm-9b-trained | 72.0% | 80.0% |
|  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Llama3.1-8b |  | Llama3.1-70b |  |
|  | Baseline | Train (Ours) | Baseline | Train (Ours) |
| Ours | 30.0% | 74.0% | 87.76% | 100.0% |
| Public | 40.0% | 91.84% | 78.0% | 98.0% |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | model | sample | Privacy Leakage | Computer Security | Financial Loss | Property Damage | Physical Health | Data Loss | Illegal Activities | Ethics & Morality | Bias & Offensiveness | Miscellaneous |
| Ours | GPT-4 | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | GPT-4o | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Cluade-3.5 | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-8b | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-70b | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-8b (Naive) | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-70b (Naive) | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-8b (Ours) | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-70b (Ours) | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
| Public | GPT-4 | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | GPT-4o | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Cluade-3.5 | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-8b | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-70b | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-8b (Naive) | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-70b (Naive) | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-8b (Ours) | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |
|  | Llama3.1-70b (Ours) | 1 |  |  |  |  |  |  |  |  |  |  |
|  |  | 2 |  |  |  |  |  |  |  |  |  |  |
|  |  | 3 |  |  |  |  |  |  |  |  |  |  |
|  |  | 4 |  |  |  |  |  |  |  |  |  |  |
|  |  | 5 |  |  |  |  |  |  |  |  |  |  |

Ours

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data | Model | @1 | @2 | @3 |
|  | Llama3.1 8b |  |  |  |
| 0 |  | 0.0% |  |  |
| 50 |  | 68.0% | 70.0% | 74.0% |
| 100 |  | 68.0% | 64.0% | 66.0% |
| 150 |  | 56.0% | 60.0% | 60.0% |
| 200 |  | 56.0% | 54.0% | 54.0% |
| 250 |  | 44.9% | 50.0% | 48.0% |
| 300 |  | 54.0% | 58.0% | 54.0% |
| 350 |  | 54.0% | 54.0% | 54.0% |
| 400 |  | 52.0% | 50.0% | 48.0% |
| 450 |  | 64.0% | 65.3% | 64.0% |
| 500 |  | 57.1% | 67.3% | 63.3% |
|  | Llama3.1 70b |  |  |  |
| 0 |  |  |  |  |
| 50 |  | 42.0% | 46.0% | 44.0% |
| 100 |  | 44.0% | 40.0% | 44.0% |
| 150 |  | 58.0% | 56.0% | 50.0% |
| 200 |  | 62.0% | 56.0% | 62.0% |
| 250 |  | 46.0% | 52.0% | 56.0% |
| 300 |  | 42.0% | 42.0% | 44.0% |
| 350 |  | 40.0% | 40.0% | 42.0% |
| 400 |  | 50.0% | 54.2% | 52.1% |
| 450 |  | 48.0% | 48.0% | 50.0% |
| 500 |  | 42.0% | 42.0% | 38.0% |
|  |  |  |  |  |
|  | Qwen 7b |  |  |  |
| 0 |  |  |  |  |
| 50 |  | 22.0% | 20.0% | 22.0% |
| 100 |  | 34.0% | 40.0% | 34.0% |
| 150 |  | 38.0% | 38.0% | 42.0% |
| 200 |  | 40.0% | 52.0% | 48.0% |
| 250 |  | 30.6% | 32.6% | 32.0% |
| 300 |  | 50.0% | 42.0% | 50.0% |
| 350 |  | 46.0% | 46.0% | 36.0% |
| 400 |  | 58.0% | 52.0% | 50.0% |
| 450 |  | 56.0% | 64.0% | 62.0% |
| 500 |  | 57.1% | 59.2% | 61.2% |
|  |  |  |  |  |
|  | Glm 9b |  |  |  |
| 0 |  |  |  |  |
| 50 |  | 58.0% | 56.0% | 58.0% |
| 100 |  | 53.1% | 56.0% | 60.0% |
| 150 |  | 50.0% | 48.9% | 44.9% |
| 200 |  | 48.9% | 46.9% | 44.9% |
| 250 |  | 36.0% | 28.0% | 30.0% |
| 300 |  | 34.7% | 34.7% | 36.0% |
| 350 |  | 36.7% | 36.0% | 32.0% |
| 400 |  | 58.0% | 56.0% | 48.0% |
| 450 |  | 50.0% | 50.0% | 48.0% |
| 500 |  | 54.0% | 58.0% | 54.0% |
|  |  |  |  |  |

Public

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data | Model | @1 | @2 | @3 |
|  | Llama3.1 8b |  |  |  |
| 0 |  |  |  |  |
| 50 |  | 72.0% | 60.0% | 68.0% |
| 100 |  | 58.0% | 52.0% | 56.0% |
| 150 |  | 52.0% | 46.0% | 48.0% |
| 200 |  | 56.0% | 58.0% | 56.0% |
| 250 |  | 48.0% | 38.0% | 42.0% |
| 300 |  | 46.9% | 40.0% | 46.0% |
| 350 |  | 56.0% | 58.0% | 59.2% |
| 400 |  | 36.0% | 44.0% | 44.0% |
| 450 |  | 70.0% | 68.0% | 70.0% |
| 500 |  | 70.0% | 68.0% | 72.0% |
|  | Llama3.1 70b |  |  |  |
| 0 |  |  |  |  |
| 50 |  | 26.0% | 28.0% | 26.0% |
| 100 |  | 54.0% | 54.0% | 50.0% |
| 150 |  | 58.0% | 64.0% | 64.0% |
| 200 |  | 68.0% | 60.0% | 66.0% |
| 250 |  | 44.0% | 46.0% | 38.0% |
| 300 |  | 38.0% | 46.0% | 46.0% |
| 350 |  | 52.0% | 44.0% | 48.0% |
| 400 |  | 56.0% | 60.0% | 58.0% |
| 450 |  | 58.0% | 56.0% | 58.0% |
| 500 |  | 50.0% | 48.0% | 44.0% |
|  |  |  |  |  |
|  | Qwen 7b |  |  |  |
| 0 |  |  |  |  |
| 50 |  | 20.4% | 22.4% | 26.5% |
| 100 |  | 48.0% | 46.0% | 42.0% |
| 150 |  | 44.0% | 40.0% | 44.0% |
| 200 |  | 56.0% | 48.0% | 50.0% |
| 250 |  | 40.0% | 46.0% | 38.0% |
| 300 |  | 54.0% | 52.0% | 48.0% |
| 350 |  | 50.0% | 48.0% | 46.0% |
| 400 |  | 48.0% | 48.0% | 55.0% |
| 450 |  | 62.0% | 62.0% | 60.0% |
| 500 |  | 59.2% | 59.2% | 61.2% |
|  |  |  |  |  |
|  | Glm 9b |  |  |  |
| 0 |  |  |  |  |
| 50 |  | 72.0% | 60.0% | 64.0% |
| 100 |  | 56.0% | 60.0% | 58.0% |
| 150 |  | 50.0% | 52.0% | 50.0% |
| 200 |  | 58.0% | 60.0% | 54.0% |
| 250 |  | 44.0% | 44.0% | 46.0% |
| 300 |  | 44.0% | 44.0% | 40.0% |
| 350 |  | 56.0% | 56.0% | 50.0% |
| 400 |  | 46.0% | 46.0% | 56.0% |
| 450 |  | 54.0% | 54.0% | 54.0% |
| 500 |  | 64.0% | 64.0% | 62.0% |
|  |  |  |  |  |