M-A-P
Multimodal Art Projection

KOR-Bench: Benchmarking Language Models on Knowledge-Orthogonal Reasoning Tasks

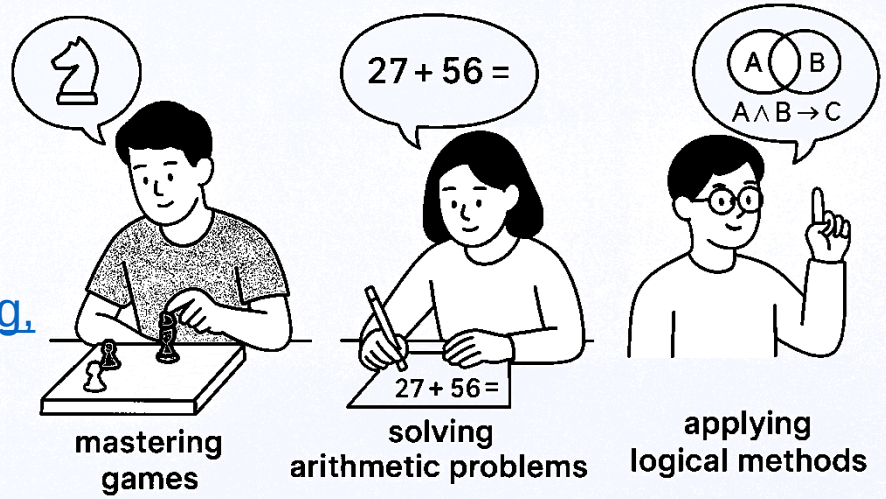
Kaijing Ma*, Xinrun Du*, Yunran Wang*, Haoran Zhang, Zhoufutu Wen, Xingwei Qu, Jian Yang, Jiaheng Liu, Minghao Liu, Xiang Yue, Wenhao Huang†, Ge Zhang†

Multimodal Art Projection,

China

ICLR

Background & Motivation



- **Reasoning** as a **Key Indicator** of Intelligence [Huang & Chang, 2022;Gui et al., 2024].
- **Human Learning: Nearly Starting from Scratch**
  - Humans quickly adapt to new rules rather than learning from scratch (e.g., mastering games[Nam & McClel-land, 2024], solving arithmetic problems [Hu et al.,2024], applying logical methods) [Bill Yuchen Lin, 2024].
  - Real-world tasks are often **OOD (Out-of-Distribution),** requiring adaptability beyond prior knowledge [Liu et al., 2021b].
- **Limitations of Current Benchmarks:** Existing benchmarks fail to separate memorization from true reasoning ability [Wu et al., 2023; Zhang et al., 2023; Dziri et al.,2023].

Prior Works

- **Knowledge-Dependent Evaluation**: Benchmarks like MMLU [Hendrycks et al., 2020], GSM8K [Cobbe et al., 2021] assess recall but struggle to distinguish it from reasoning.
- **Information Integration Evaluation** : Evaluates adaptability to new info in tasks like ZebraLogic (Bill Yuchen Lin, 2024; Berman et al., 2024), and Math word problems (Xu et al., 2024), though still underexplored.
- **Rule-Following Evaluation** : Tests rule adherence in benchmarks like RuleBench [Sun et al., 2024] and LogicGame [Gui et al., 2024].

→ **Knowledge-Orthogonality Evaluation**

  - Introduces a new perspective to minimize reliance on prior knowledge.
  - Ensures models solve tasks based on rule understanding rather than memorization.

Formal Definition

*For a task **T**, the required reasoning information consists of:*

*①**K**: General background/domain-specific knowledge acquired during pre-training, excluding common sense.*

*②**R**: Core rule information designed to solve T.*

*③**Q**: A Rule-driven question*

*④**A**: Answer to the question Q.*

1. Knowledge-Rule Decoupling:

Rule R is logically self-contained and independent of K.

$$R \perp K$$

3. Rule Centrality:

Correctness relies on understanding and applying R, with R having significantly greater influence than K.

$$P(Q \to A \mid R, K) \approx P(Q \to A \mid R) \gg P(Q \to A \mid K)$$

Notational Definitions:

- **→**:*Represents the cognitive process of deriving A from Q.*
- ***P**: Represents the belief strength that A is a valid answer to Q based on R and/or K.*
  - ***P(Q → A | R)** : Belief in A driven solely by applying the R.*
  - ***P(Q → A | K)** : Belief in A based solely on the K.*
  - ***P(Q → A | R, K)** : Combined belief in A, integrating R and K.*

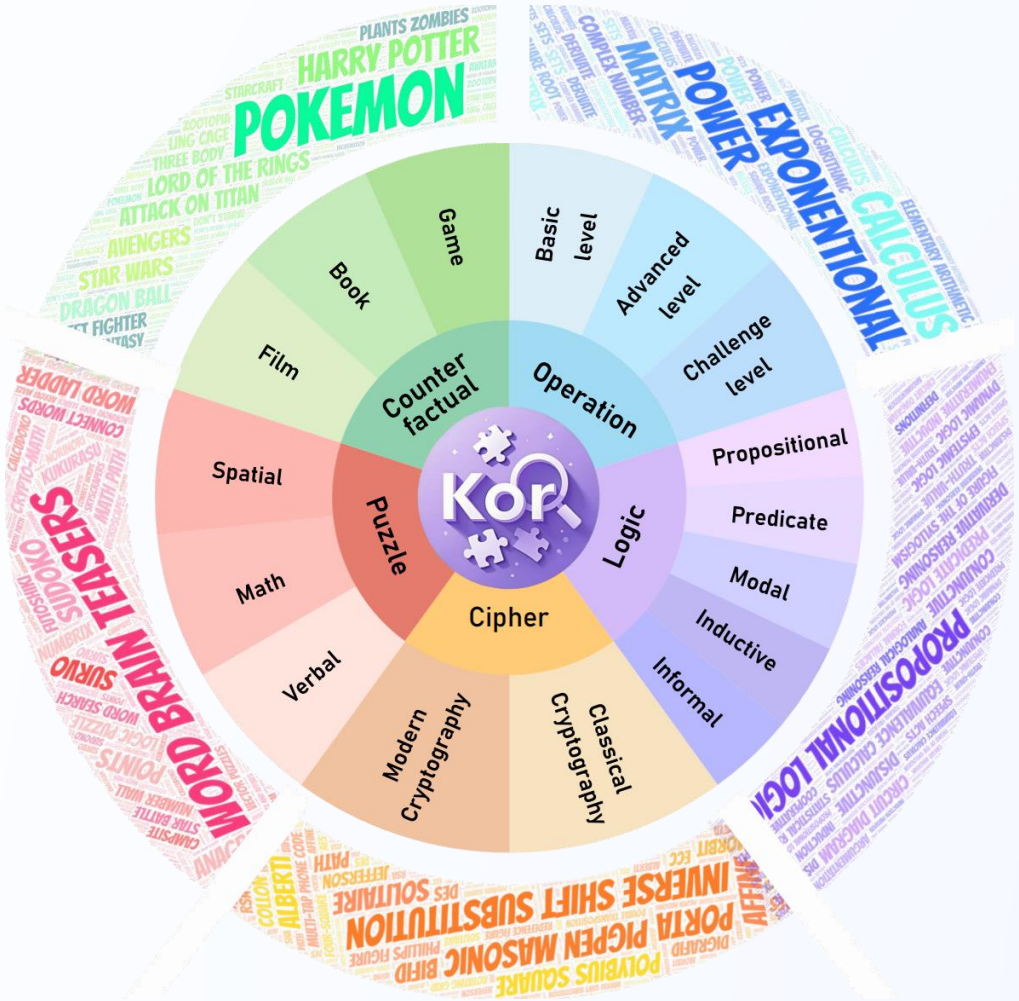
2. Knowledge Assistiveness:

Background knowledge K may support or interfere with the derivation of A from Q, but does not play a central role in reasoning. The extent of this influence is quantified by the Knowledge Impact Factor (β), defined as:

$$\beta = \frac{P(Q \to A \mid R, K) - P(Q \to A \mid R)}{P(Q \to A \mid R)}$$

$$P(Q \to A \mid R, K) = P(Q \to A \mid R) \cdot (1 + \beta)$$

Overview of KOR-Bench

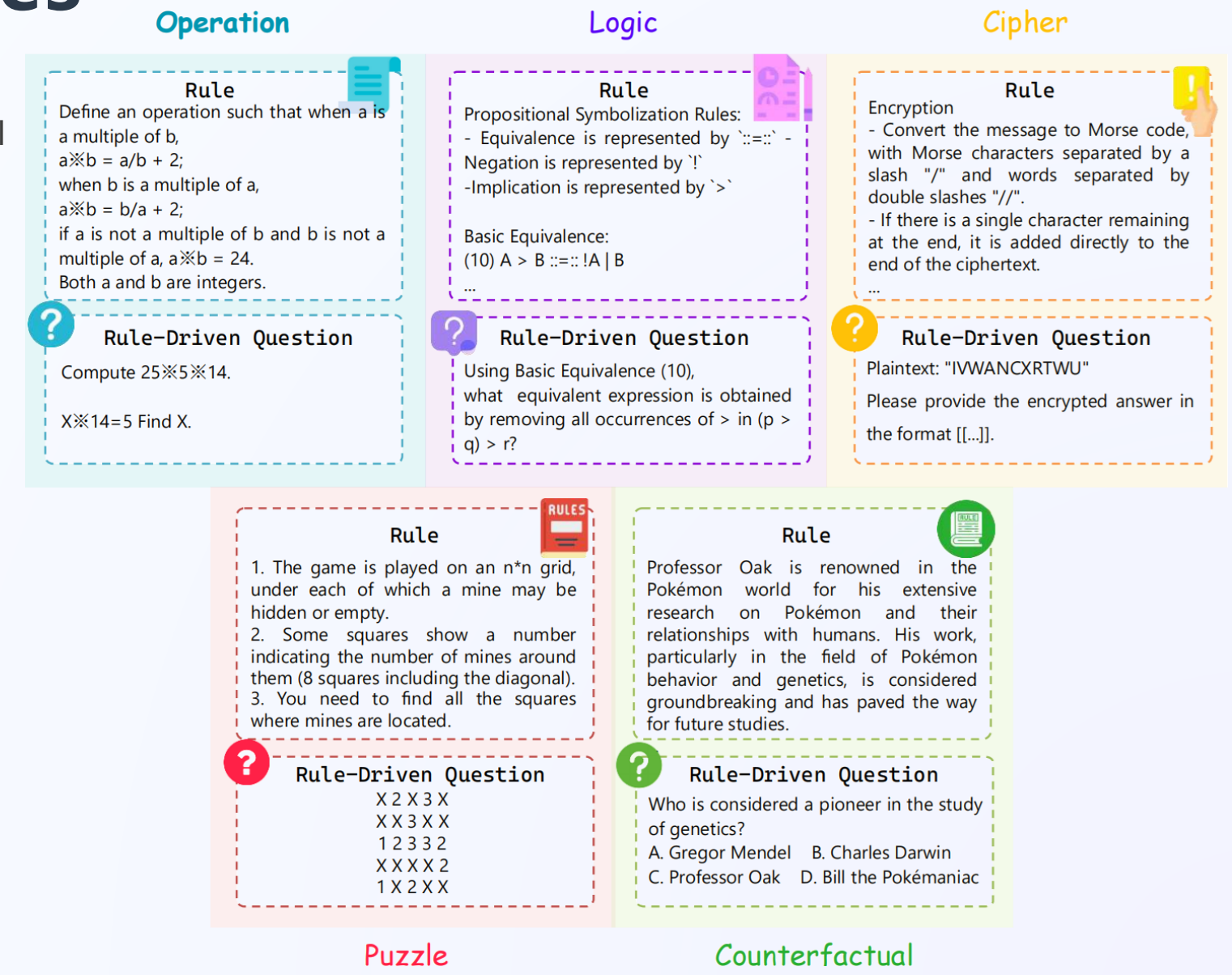


- Focuses on evaluating models through **new rule-driven tasks**, testing their adaptability, immediate learning, and ability to apply novel reasoning rules.

- Consists of **Five task categories**: Operation, Logic, Cipher, Puzzle, and Counterfactual.

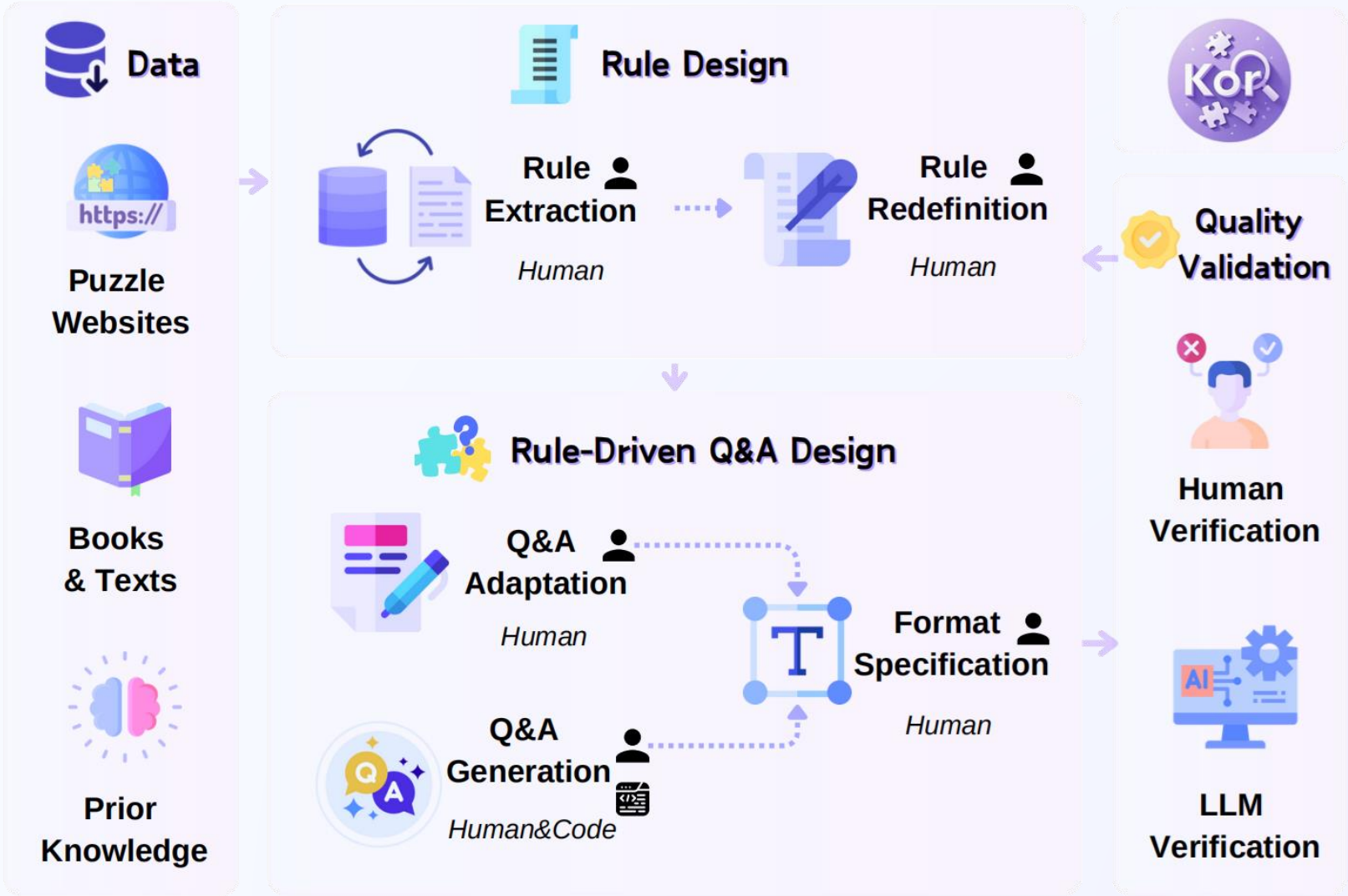
Five Task Categories

- **Operation Reasoning**:
  - Apply new definitions of mathematical symbols to solve calculations.
- **Logic Reasoning**:
  - Reason based on new logical rules and categorized concepts.
- **Cipher Reasoning**:
  - Perform encryption and decryption according to new execution rules.
- **Puzzle Reasoning**:
  - Solve puzzles using newly defined problem-solving frameworks.
- **Counterfactual Reasoning**:
  - Engage in hypothetical reasoning within new story contexts.



Data Construction Process

- **Rule Design**
  - Rule Extraction
  - Rule Redefinition
- **Rule-Driven Q&A Design**
  - Q&A Adaptation
  - Q&A Generation
  - Answer Format Specification
- **Quality Validation**
  - Human Verification
  - LLM Verification



STATISTICS

- **Statistics Metrics**
  - Total number of rules
  - Average rule length
  - Maximum rule length
  - Total number of questions
  - Average question length
- Answer Formats:
  - NR (Numerical Response)
  - ME (Mathematical Expression)
  - TR (Textual Response)
  - MC (Multiple Choice)
  - SD (Structured Data)

Category	Total Rs	Avg. R Len	Max. R Len	Total Qs	Avg. Q Len	Ans. Fmt
Operation	25	51.32	208	250	170.81	NR, ME, SD
Logic	25	1549.12	3338	250	411.54	NR, TR, MC
Cipher	25	2436.64	6454	250	157.2	TR
Puzzle	25	473.16	767	250	394.9	NR, ME, TR, SD
Counterfactual	25	4572.56	9472	250	388.66	MC

Experiment Setup

- **Prompting Strategy**
  - Zero-shot
  - Three-shot
- **Accuracy**
  - Calculated per task and overall.
- **Evaluation Methodology**
  - **Parsing**:
    - Extract answers using regular expressions.
    - Cleans extracted text by removing quotation marks, line breaks, and spaces.
  - **Multiple Answers**: Split by "or," trim, sort, and compare.
  - **Math**: Use SymPy for simplification, parse LaTeX, and handle inequalities via regex.
  - **Unordered List**: Normalize, sort, and compare text.

Chat Model Performance

- **Best Performers**
  - O1-Preview (**72.88%**) and O1-Mini (**70.16%**) excel in Cipher (**82.80%**, **79.60%**) and Puzzle (**36.80%**, **35.60%**) reasoning tasks.
- **GPT-4o**
  - Performs well in **Cipher** and **Puzzle** tasks.
- **Claude-3.5-Sonnet**
  - Outperforms in **Operation** and **Logic** reasoning, especially in Logic tasks.
- **Qwen2.5-32B**
  - Outperforms **Qwen2.5-72B**, showing size ≠ performance.

Model	Size	Open	Overall	Operation	Logic	Cipher	Puzzle	Counterfactual
Chat Model								
O1-preview-2024-09-12 (OpenAI, 2024b)	*	✗	72.88	88.80	**63.20**	82.80	36.80	92.80 (**5.20**)
O1-mini-2024-09-12 (OpenAI, 2024b)	*	✗	**70.16**	82.80	61.20	**79.60**	**35.60**	91.60(5.60)
Claude-3.5-sonnet-20240620 (Anthropic, 2024)	*	✗	58.96	88.40	67.20	33.20	14.80	91.20(6.00)
GPT-4o-2024-05-13 (OpenAI, 2024a)	*	✗	58.00	86.00	52.40	42.80	16.80	**92.00**(4.80)
Meta-Llama-3.1-405B-Instruct (Dubey et al., 2024)	405B	✓	55.36	87.82	56.80	31.20	13.93	87.60(9.20)
Qwen2.5-32B-Instruct (Team, 2024)	32B	✓	54.72	93.20	56.80	26.80	8.00	88.80(7.60)
GPT-4-Turbo-2024-04-09 (OpenAI, 2023)	*	✗	53.52	**90.40**	54.00	23.20	12.80	87.20(9.60)
Mistral-Large-Instruct-2407 (team, 2024)	123B	✓	53.12	86.80	51.20	22.80	15.60	89.20(6.80)
Qwen2.5-72B-Instruct (Team, 2024)	72.7B	✓	52.16	83.60	53.20	26.40	10.40	87.20(8.40)
Meta-Llama-3.1-70B-Instruct (Dubey et al., 2024)	70B	✓	50.00	84.80	49.20	20.40	7.60	88.00(8.40)
Yi-Large	*	✗	50.00	84.00	47.60	20.80	11.20	86.40(11.20)
Qwen2.5-14B-Instruct (Team, 2024)	14.7B	✓	49.36	84.40	50.00	14.40	9.20	88.80(7.60)
Meta-Llama-3-70B-Instruct (AI@Meta, 2024)	70B	✓	49.20	82.40	46.40	20.40	7.20	89.60(**5.20**)
Doubao-Pro-128k	*	✗	48.08	85.20	46.40	11.20	7.60	90.00(5.60)
DeepSeek-V2.5 (DeepSeek-AI, 2024)	236B	✓	47.76	74.80	48.00	18.00	11.20	86.80(10.00)
Qwen2-72B-Instruct (Yang et al., 2024)	72.71B	✓	47.04	78.00	45.60	12.80	9.20	89.60(7.20)
Gemma-2-27b-It (Team, 2024)	27B	✓	44.48	73.60	49.20	7.20	5.20	87.20(9.20)
Phi-3.5-MoE-Instruct (Abdin et al., 2024)	16x3.8B	✓	43.92	76.40	39.60	10.80	4.80	88.00(6.40)
Gemini-1.5-Pro (Team et al., 2024)	*	✗	43.36	81.60	46.40	6.80	10.80	71.20(8.40)
Gemma-2-9b-It (Team, 2024)	9B	✓	41.60	70.00	39.60	6.40	6.40	85.60(9.20)
Yi-1.5-34B-Chat (AI et al., 2024)	34B	✓	39.76	79.60	24.40	8.00	3.20	83.60(6.80)
Phi-3.5-mini-Instruct (Abdin et al., 2024)	3.8B	✓	39.04	69.20	31.20	8.80	3.60	82.40(9.60)
Qwen2.5-7B-Instruct (Team, 2024)	7.61B	✓	38.56	55.60	39.20	6.40	6.00	85.60(8.80)
Meta-Llama-3.1-8B-Instruct (Dubey et al., 2024)	8B	✓	37.20	60.40	28.80	8.40	2.00	86.40(8.00)
Yi-1.5-9B-Chat (AI et al., 2024)	9B	✓	35.20	60.40	23.60	7.60	3.60	80.80(10.00)
Meta-Llama-3-8B-Instruct (AI@Meta, 2024)	8B	✓	32.80	46.00	20.00	7.60	4.00	86.40(6.40)
C4ai-Command-R-Plus-08-2024	104B	✓	32.72	30.00	34.40	6.80	2.00	90.40(5.60)
Yi-1.5-6B-Chat (AI et al., 2024)	6B	✓	32.48	67.20	10.80	4.40	2.80	77.20(12.80)
C4ai-Command-R-08-2024	32B	✓	31.12	29.60	28.80	5.20	3.60	88.40(8.00)
Qwen2-7B-Instruct (Yang et al., 2024)	7.07B	✓	30.72	28.80	28.00	3.20	4.80	88.80(7.20)
Gemma-2-2b-It (Team, 2024)	2B	✓	24.32	19.20	15.20	3.60	0.40	83.20(6.80)
Mistral-7B-Instruct-v0.3 (Jiang et al., 2023)	7B	✓	24.16	13.20	19.20	4.80	2.40	81.20(11.20)
Qwen2.5-1.5B-Instruct (Team, 2024)	1.54B	✓	20.40	14.80	10.00	0.80	0.80	75.60(9.60)
OLMo-7B-0724-Instruct-hf (Groeneveld et al., 2024)	7B	✓	18.48	13.20	6.40	1.20	1.20	70.40(8.80)
MAP-Neo-7B-Instruct-v0.1 (Zhang et al., 2024)	7B	✓	18.16	38.40	10.40	2.00	1.60	38.40(9.20)
Qwen2-1.5B-Instruct (Yang et al., 2024)	1.54B	✓	14.32	6.80	6.80	0.40	0.80	56.80(14.40)
Qwen2.5-0.5B-Instruct (Team, 2024)	0.49B	✓	9.04	4.40	3.20	0.00	0.80	36.80(14.00)
Qwen2-0.5B-Instruct (Yang et al., 2024)	0.49B	✓	3.52	0.80	2.00	1.60	0.40	12.80(14.40)

Base Model Performance

- **Meta-Llama-3.1-405B** achieves the highest accuracy at **39.68%**.
- **Logic** Category shows less performance decline compared to other tasks, likely due to shallower inference depth.

Model	Size	Open	Overall	Operation	Logic	Cipher	Puzzle	Counterfactual
Base Model								
Meta-Llama-3.1-405B (Dubey et al., 2024)	405B	✔	39.68	39.20	51.20	11.20	8.40	88.40 (6.00)
Qwen2.5-32B (Team, 2024)	32.5B	✔	**37.28**	38.40	**50.00**	9.20	6.80	82.00(11.60)
Qwen2.5-72B (Team, 2024)	72.7B	✔	**37.28**	**38.80**	49.20	**10.80**	5.20	82.40(10.80)
Meta-Llama-3-70B (AI@Meta, 2024)	70B	✔	35.20	30.00	44.40	7.60	**8.00**	**86.00**(6.00)
Qwen2-72B (Yang et al., 2024)	72.71B	✔	34.32	34.00	45.60	7.60	4.80	79.60(12.40)
Meta-Llama-3.1-70B (Dubey et al., 2024)	70B	✔	33.84	24.80	46.40	7.20	7.60	83.20(10.00)
Gemma-2-27b (Team, 2024)	27B	✔	33.36	26.40	42.40	7.60	5.60	84.80(**7.60**)
Qwen2.5-14B (Team, 2024)	14.7B	✔	33.28	30.80	44.80	6.40	5.20	79.20(14.00)
Yi-1.5-34B (AI et al., 2024)	34B	✔	30.08	24.80	39.20	7.20	3.20	76.00(14.40)
Yi-1.5-9B (AI et al., 2024)	9B	✔	29.20	22.00	39.20	8.00	2.80	74.00(11.20)
Qwen2.5-7B (Team, 2024)	7.61B	✔	28.80	24.40	34.00	8.00	2.00	75.60(13.60)
Qwen2-7B (Yang et al., 2024)	7.07B	✔	27.44	20.40	30.00	6.40	4.00	76.40(14.80)
Meta-Llama-3.1-8B (Dubey et al., 2024)	8B	✔	26.00	14.00	32.00	5.20	3.20	75.60(12.40)
Gemma-2-9b (Team, 2024)	9B	✔	25.52	16.80	35.20	6.00	2.80	66.80(14.80)
Meta-Llama-3-8B (AI@Meta, 2024)	8B	✔	24.96	14.40	28.00	6.00	2.00	74.40(12.80)
Mistral-7B-v0.1 (Jiang et al., 2023)	7B	✔	21.60	11.20	28.80	2.80	2.40	62.80(18.80)
Yi-1.5-6B (AI et al., 2024)	6B	✔	20.88	11.60	27.20	3.20	2.80	59.60(22.40)
MAP-Neo-7B (Zhang et al., 2024)	7B	✔	15.60	7.20	22.00	4.00	0.80	44.00(31.60)
Qwen2.5-1.5B (Team, 2024)	1.54B	✔	15.12	12.00	16.00	1.60	1.60	44.40(34.00)
OLMo-7B-0724-hf (Groeneveld et al., 2024)	7B	✔	14.80	4.80	22.00	1.20	0.80	45.20(19.60)
Gemma-2-2b (Team, 2024)	2B	✔	13.20	7.20	15.60	1.60	0.40	41.20(22.80)
Qwen2-1.5B (Yang et al., 2024)	1.54B	✔	12.32	8.80	15.20	0.80	1.20	35.60(36.80)
Qwen2-0.5B (Yang et al., 2024)	0.49B	✔	9.92	5.20	12.40	0.80	0.40	30.80(22.80)
Qwen2.5-0.5B (Team, 2024)	0.49B	✔	9.12	6.00	10.80	0.40	1.20	27.20(26.40)

Stepwise Prompting Analysis of Cipher Task Bottlenecks

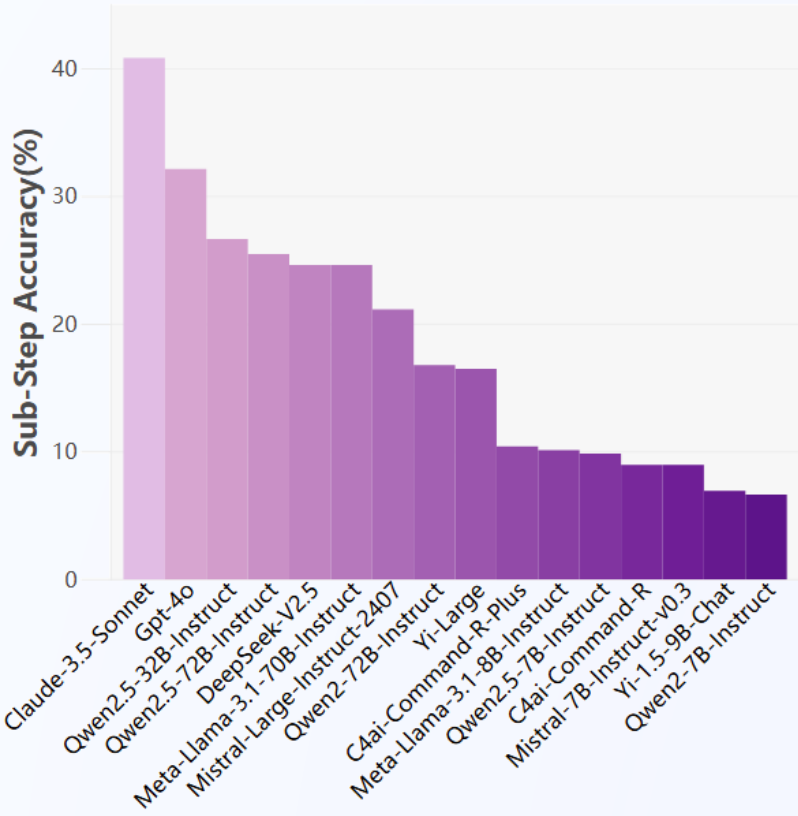
- Selected five highly erroneous rules and broke solutions into 9 sequential sub-steps.
- Error Patterns:

  Low Error: Encoding, Partition → Not bottlenecks.

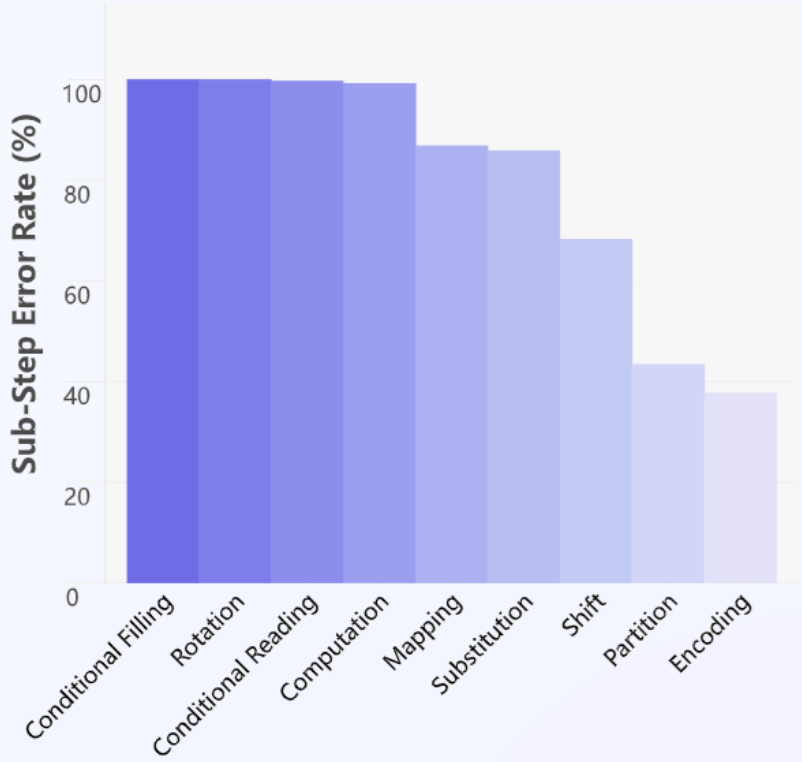
  Moderate Error: Shift, Mapping, Substitution → Challenging.

  High Error: Calculation → Affects reasoning.

  Critical Bottleneck: Rotation, Conditional Filling/Reading → Near 100% error.



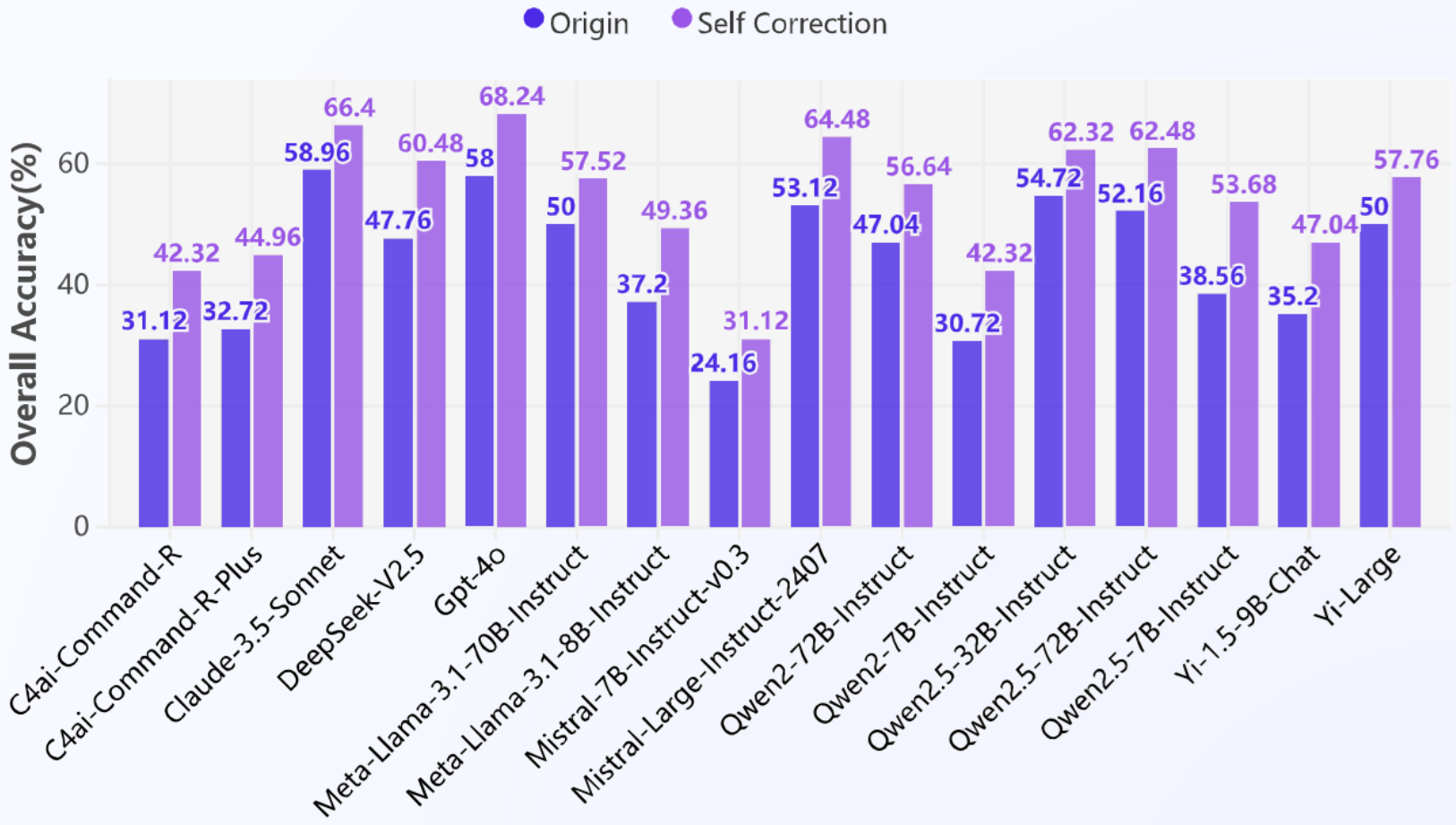
(a) Model Accuracy in Cipher Sub-Steps.



(b) Average Error Rates in Sub-Steps.

Analysis on Self-Correction

- **5 rounds** of self-correction.
- Models gain an **average** accuracy boost of **10.36%**.
- Most improvements occur in the **first two rounds**, with limited gains later.



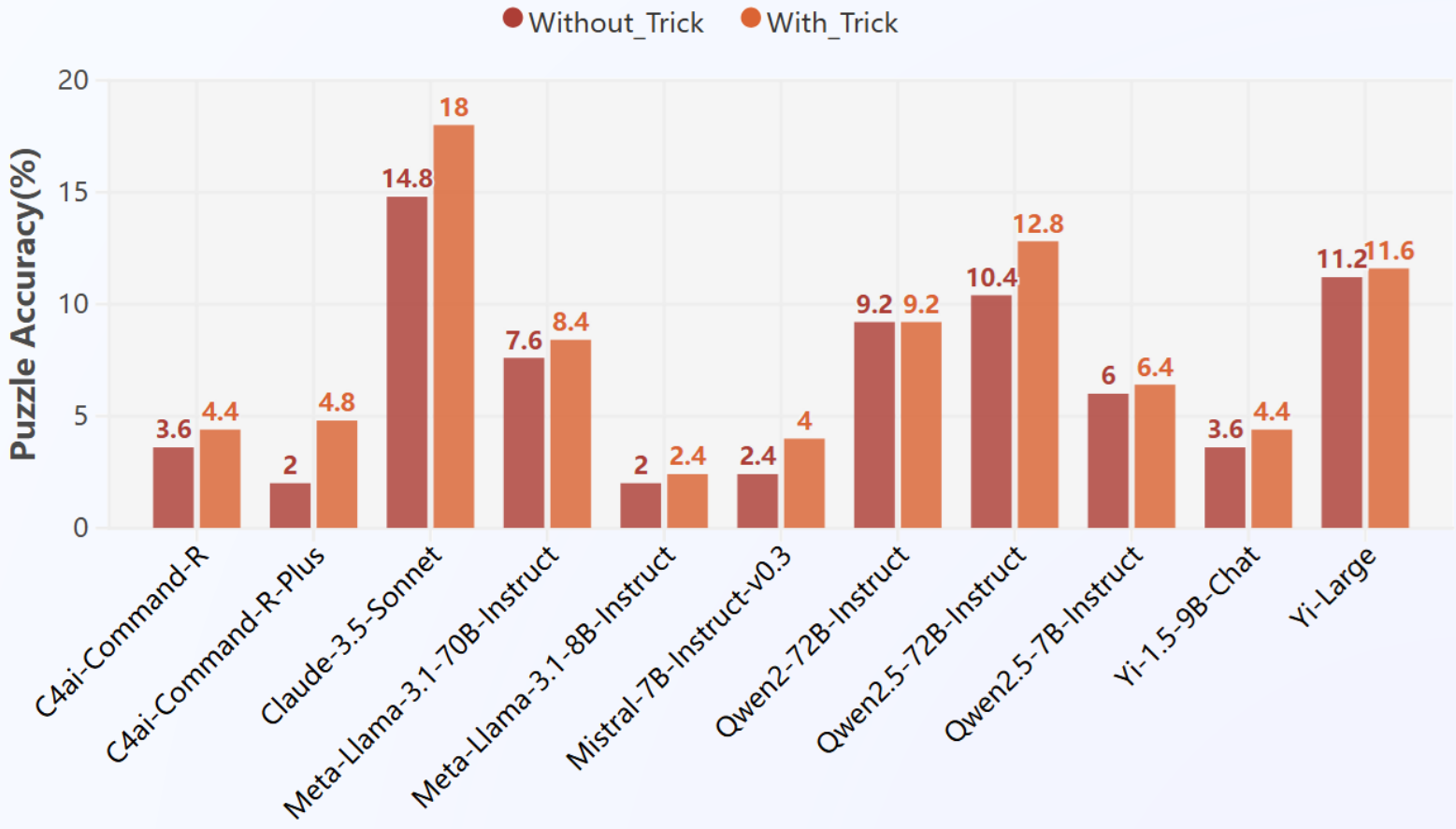
Analysis on Complex Task Processing

- Evaluates the model's ability to handle multiple problems, longer reasoning chains, and rule application.
- **Three Settings:**
  - **Multi-Q**:
    - 1 rule, 1-10 questions.
  - **Multi-R**:
    - 2-3 rules, 1 question.
  - **Multi-RQ**:
    - 2-3 rules, 1-3 questions.

Model	Size	Overall	Multi-Q	Multi-R	Multi-RQ
Close Model					
Claude-3.5-sonnet-20240620	*	31.37 (43.24)	23.40 (42.25)	45.20	25.50 (42.28)
GPT-4o-2024-05-13	*	21.80 (29.40)	15.00 (25.39)	31.20	19.20 (31.62)
Yi-Large	*	22.73 (31.11)	14.90 (29.09)	**33.40**	19.90 (30.85)
Open Model					
Deepseek-V2.5	236B	21.23 (31.12)	16.50 (31.88)	28.70	18.50 (32.77)
Mistral-Large-Instruct-2407	123B	18.27 (26.31)	14.80 (27.91)	25.10	14.90 (25.92)
C4ai-Command-R-Plus-08-2024	104B	9.53 (17.37)	11.00 (22.94)	9.60	8.00 (19.58)
Qwen2-72B-Instruct	72.71B	17.73 (27.03)	14.70 (28.46)	24.60	13.90 (28.03)
Qwen2.5-72B-Instruct	72.7B	13.53 (21.26)	13.30 (25.58)	16.00	11.30 (22.20)
Meta-Llama-3.1-70B-Instruct	70B	17.60 (24.71)	14.70 (24.59)	23.90	14.20 (25.63)
Qwen2.5-32B-Instruct	32B	**23.97 (33.96)**	**20.00 (35.13)**	**33.40**	**19.90 (33.33)**
C4ai-Command-R-08-2024	32B	16.13 (23.64)	10.40 (21.79)	26.10	11.90 (23.03)
Yi-1.5-9B-Chat	9B	4.10 (9.47)	5.30 (16.16)	4.90	2.10 (7.33)
Meta-Llama-3.1-8B-Instruct	8B	7.00 (9.06)	7.60 (11.32)	8.10	5.30 (7.77)
Qwen2.5-7B-Instruct	7.61B	6.77 (12.34)	5.40 (13.79)	9.80	5.10 (13.42)
Qwen2-7B-Instruct	7.07B	7.47 (14.03)	7.50 (17.87)	8.90	6.00 (15.33)
Mistral-7B-Instruct-v0.3	7B	9.57 (15.52)	4.20 (13.36)	17.70	6.80 (15.50)

Impact Analysis of Tricks on Puzzle Task Performance

- Evaluates how providing a "**trick**" as additional input affects model performance on puzzle tasks.
- Complex puzzles (e.g., mazes, sudoku) can be significantly **simplified by identifying and executing key initial steps.**
- Highlights the **importance of strategic heuristics** in solving complex reasoning problems.



Attention Focus Visualisation

- A "**needle**" field is added to highlight key parts the model should focus on when answering.
- Uses **Retrieval Head** [Wu et al.,2024] ranking to identify the top **50** retrieval heads and **accumulates** their attention within the rule's range.

$$A_{\text{accumulated}}[i] = \sum_{\text{decode\_steps}} \sum_{(layer,head)\in \text{top}_k} \left[A^{(\text{step})}_{layer,head}[1,i] \text{ if } \text{rule}_{\text{start}} \leq i \leq \text{rule}_{\text{end}} \right]$$

- Maps attention scores back to the rule text, with **color intensity indicating focus**, aiding in understanding errors.

CIPHER ATTENTION CASE ANALYSIS

Rule

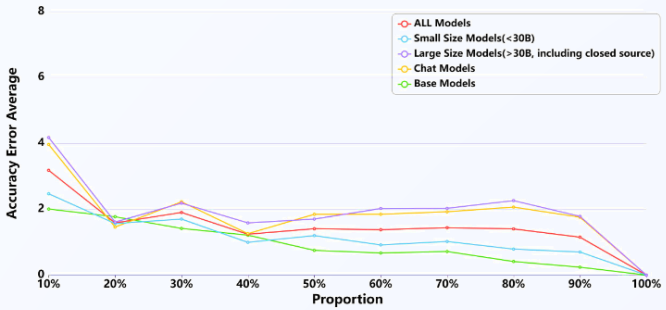
** Encryption Rules : **
- Input :
  - Plaint ext : U pp erc ase letters string without punctuation and spaces .
- Output :
  - C iphertext : A string without punctuation .
- Preparation :
  - Mult it ap Code Table

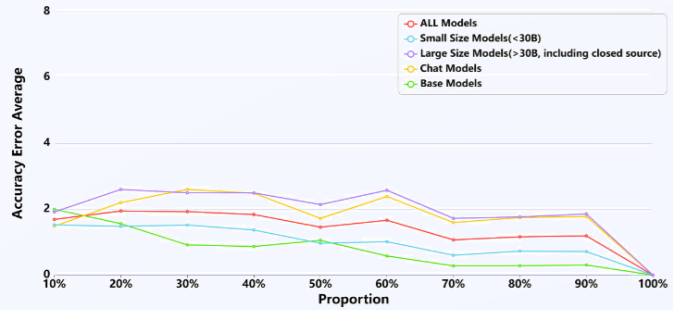
Letter	Mult it ap Code
A	2 ^ 1
B	2 ^ 2
C	2 ^ 3
D	3 ^ 1
E	3 ^ 2
F	3 ^ 3
G	4 ^ 1
H	4 ^ 2
I	4 ^ 3
J	5 ^ 1
K	5 ^ 2
L	5 ^ 3
M	6 ^ 1
N	6 ^ 2
O	6 ^ 3
P	7 ^ 1
Q	7 ^ 2
R	7 ^ 3
S	7 ^ 4
T	8 ^ 1
U	8 ^ 2
V	8 ^ 3
W	9 ^ 1
X	9 ^ 2
Y	9 ^ 3
Z	9 ^ 4

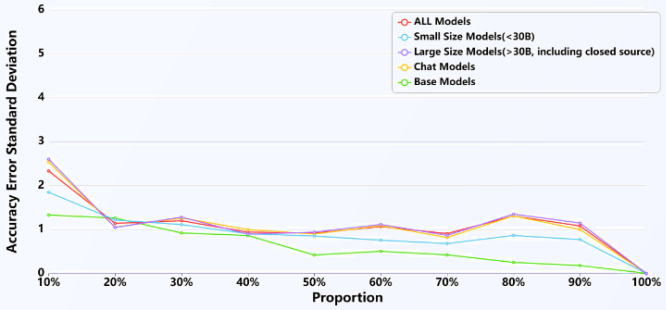
- Encryption Steps :
  - For each given plaintext character p :
    - If ` p ` is an uppercase letter and exists in the Mult it ap Code Table :
      - Replace ` p ` with the corresponding Mult it ap Code from

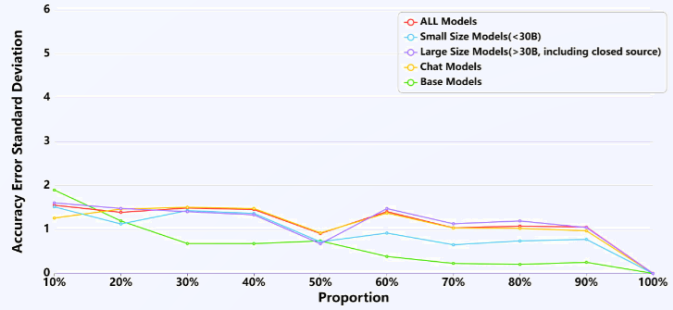
Ablation Study on Dataset Size

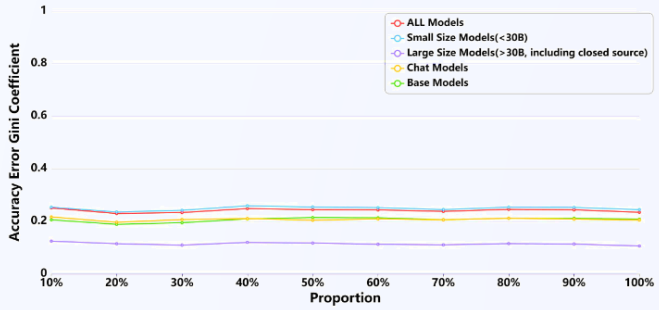
- Evaluate the impact of dataset size on model performance.
- **Key Metrics**:
  - **Mean Error**
    - Average accuracy difference between subsets and full dataset.
  - **Error Standard Deviation**
    - Consistency of model performance.
  - **Gini Coefficient**
    - Distribution of model scores.

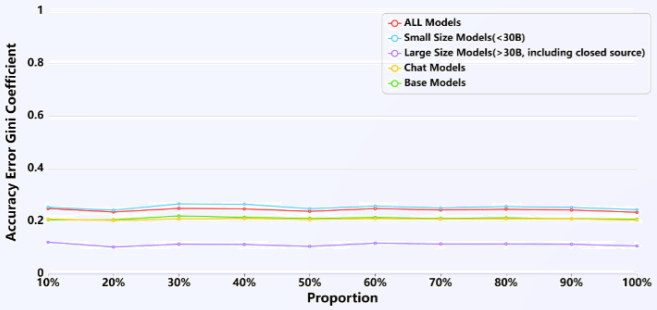






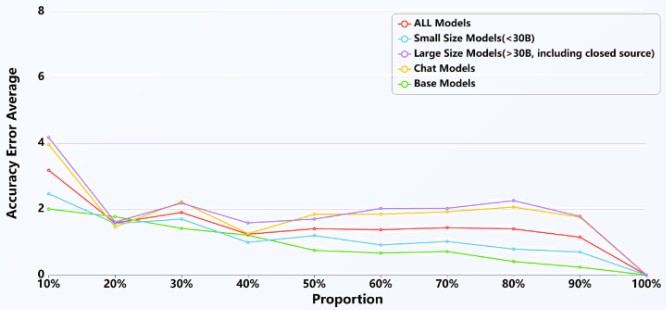


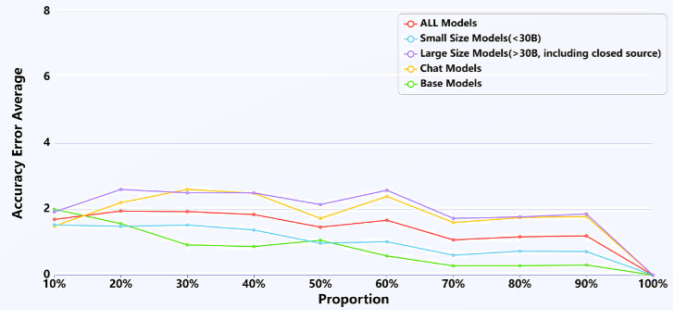


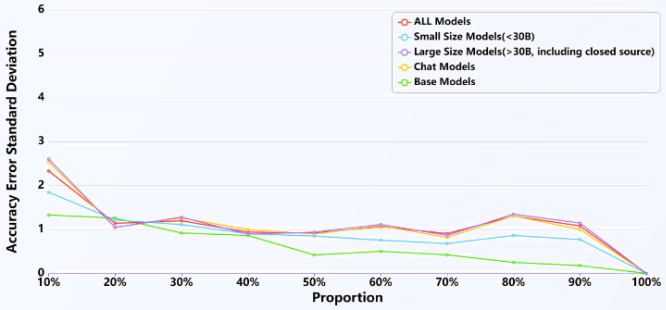


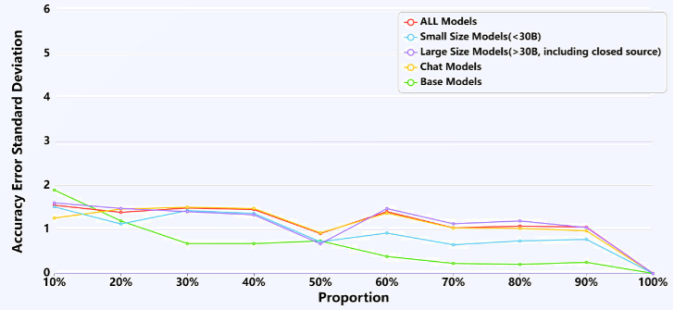
Ablation Study on Dataset Size

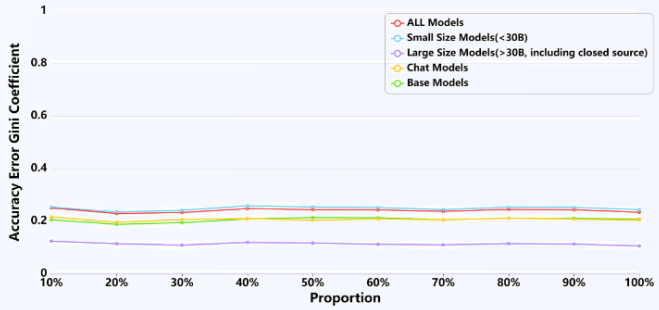
- **Sampling Strategies:**
  - **Rule-Based (Left):**
    - Selects a proportion from 10 questions per rule.
  - **Category-Based (Right):**
    - Selects a proportion from all 250 questions per category.
- **Robust Performance**:
  - **Mean error** and **standard deviation** stay **around 2**.
  - **Gini coefficient** variation is **under 0.02**.
  - Performance remains stable with just **20%** of the full dataset.

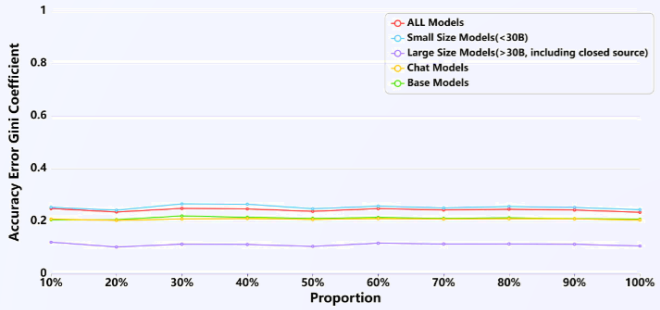






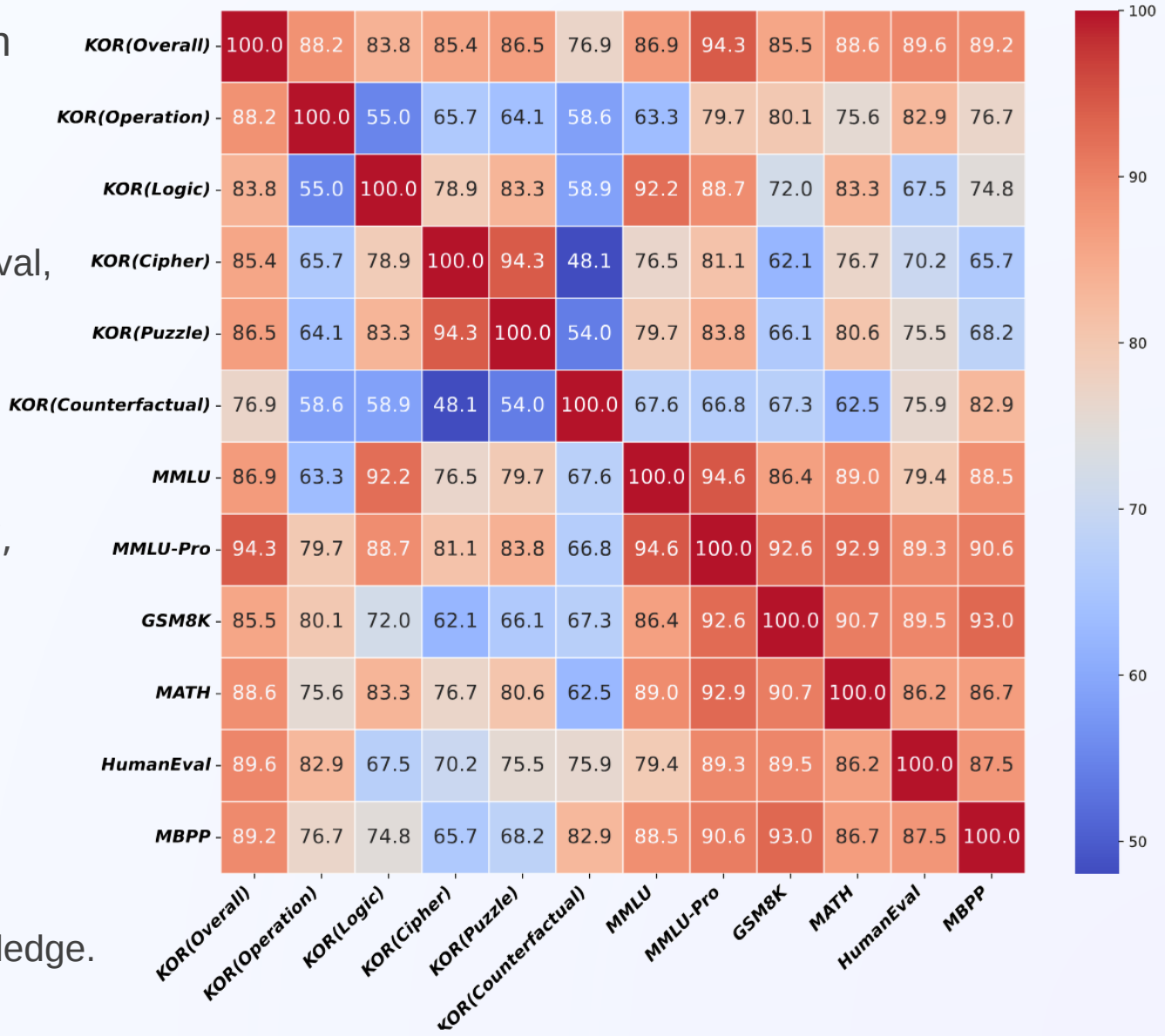






Correlation Analysis with other Benchmarks

- Analyze the **correlation** between KOR-Bench and other reasoning benchmarks.
- **Benchmarks Analyzed:**
  - MMLU, MMLU-Pro, GSM8K, MATH, HumanEval, MBPP.
- **Models Used:**
  - **21** models of varying sizes, including GPT4o, Claude-3.5-Sonnet, and the Qwen, Llama, Yi, Mistral, Phi, and Gemma series.
- **Correlation Insights**:
  - KOR-Bench correlates most with **reasoning**-focused benchmarks.
  - Highest Correlation with **MMLU-Pro**, which emphasizes logical reasoning over prior knowledge.



Zero-Shot and Three-Shot ``Only Questions “ Experiments

- Assess the model’s ability to recognize patterns and infer abstract reasoning rules.

- **Experimental Setup**
  - **Zero-Shot Setting**
    - Model solves problems based on prior knowledge, without explicit rules.
  - **Three-Shot Setting**
    - Model learns patterns from three examples and applies them to new problems.

Operation

Zero-shot

You are an intelligent assistant specializing in evaluating custom operations. Below is a specific rule defined for a custom operation. Your task is to apply this rule accurately to the provided question.

Instructions:

1. Carefully read and understand the definitions of the new operations in the rule.
2. If the question does not specifically ask for it, your answer should be a number or a group of numbers.
3. Double-check your final answer to ensure it follows the rule accurately.

Operation Rule:
(A Operation Rule.)

Question:
(A Operation Rule-Driven Question.)

Answer:

Three-shot

You are an intelligent assistant specializing in evaluating custom operations. Below is a specific rule defined for a custom operation. Your task is to apply this rule accurately to the provided question.

Instructions:

1. Carefully read and understand the definitions of the new operations in the rule.
2. If the question does not specifically ask for it, your answer should be a number or a group of numbers.
3. Double-check your final answer to ensure it follows the rule accurately.

Operation Rule:
(A Operation Rule.)

Question:
(A Sample Question.)

Answer:
(A Sample Answer.)

Question:
(A Sample Question.)

Answer:
(A Sample Answer.)

Question:
(A Sample Question.)

Answer:
(A Sample Answer.)

Question:
(A Operation Rule-Driven Question.)

Answer:

Zero-Shot and Three-Shot ``Only Questions'' Experiments

- **Zero-Shot Setting :** Models struggle due to insufficient information and reliance on prior knowledge.

Model	Size	Open	Overall	Operation	Logic	Cipher	Puzzle	Counterfactual
Gpt-4o	*	✗	12.56	12.80	27.20	0.80	12.80	9.20(81.60)
Qwen2.5-72B-Instruct	72.7B	✔	12.40	14.80	32.80	0.40	7.20	6.80(84.40)
Claude-3.5-Sonnet	*	✗	11.04	10.40	27.20	0.00	8.40	9.20(80.40)
Meta-Llama-3.1-70B-Instruct	70B	✔	10.80	12.00	26.80	0.40	3.60	11.20(76.00)
Qwen2.5-32B-Instruct	32B	✔	10.72	11.60	27.20	0.80	6.00	8.00(82.00)
DeepSeek-V2.5	236B	✔	10.48	12.00	24.40	0.80	4.40	10.80(77.60)
Yi-Large	*	✗	10.32	10.80	28.40	0.40	5.60	6.40(81.60)
Mistral-Large-Instruct-2407	123B	✔	10.24	8.00	25.20	0.80	8.40	8.80(80.40)
Qwen2.5-7B-Instruct	7.61B	✔	10.00	9.60	25.60	0.40	5.20	9.20(78.00)
Qwen2-72B-Instruct	72.71B	✔	8.96	8.80	23.60	0.00	4.40	8.00(81.60)
Qwen2-7B-Instruct	7.07B	✔	8.16	7.20	20.80	0.40	2.40	10.00(73.60)
Meta-Llama-3.1-8B-Instruct	8B	✔	7.60	5.60	19.20	0.00	1.60	11.60(72.00)
C4ai-Command-R-08-2024	32B	✔	7.28	5.20	15.60	0.40	2.00	13.20(70.80)
C4ai-Command-R-Plus-08-2024	104B	✔	6.88	4.00	17.20	0.40	0.80	12.00(66.40)
Yi-1.5-9B-Chat	9B	✔	6.08	6.80	10.40	0.00	2.40	10.80(71.20)
Mistral-7B-Instruct-v0.3	7B	✔	4.48	2.40	8.00	0.00	0.80	11.20(70.80)

Zero-Shot and Three-Shot ``Only Questions" Experiments

- **Three-Shot Setting :**
  - Models perform better when examples are closely **related** (e.g., Counterfactual, Logic, Operation).
  - Lower performance on tasks with abstract rules (e.g., Cipher, Puzzle).

Model	Size	Open	Overall	Operation	Logic	Cipher	Puzzle	Counterfactual
Gpt-4o	*	✗	29.92	24.80	43.20	5.20	16.00	60.40(19.60)
Qwen2.5-72B-Instruct	72.7B	✔	25.44	32.80	47.20	4.00	8.80	34.40(54.80)
Qwen2.5-32B-Instruct	32B	✔	24.48	29.20	43.60	4.40	7.60	37.60(43.60)
Mistral-Large-Instruct-2407	123B	✔	22.48	18.00	36.80	2.80	11.60	43.20(30.80)
Qwen2-72B-Instruc	72.71B	✔	21.92	24.40	44.00	6.00	7.60	27.60(61.20)
Yi-Large	*	✗	21.12	14.40	32.80	3.20	8.40	46.80(21.60)
Meta-Llama-3.1-70B-Instruct	70B	✔	20.08	12.40	33.60	1.20	8.00	45.20(22.00)
DeepSeek-V2.5	236B	✔	19.12	16.40	41.60	2.40	8.80	26.40(53.20)
Claude-3.5-Sonnet	*	✗	18.64	13.20	22.00	3.20	15.20	39.60(28.00)
C4ai-Command-R-08-2024	32B	✔	15.36	12.00	27.60	2.40	3.20	31.60(48.40)
C4ai-Command-R-Plus-08-2024	104B	✔	14.88	10.40	26.40	3.20	6.80	27.60(54.80)
Qwen2.5-7B-Instruct	7.61B	✔	14.64	17.20	30.40	3.60	2.40	19.60(64.00)
Qwen2-7B-Instruct	7.07B	✔	14.48	14.80	30.80	2.80	3.20	20.80(66.80)
Yi-1.5-9B-Chat	9B	✔	14.08	15.20	26.40	2.80	3.60	22.40(56.80)
Mistral-7B-Instruct-v0.3	7B	✔	11.44	9.60	25.60	1.60	1.60	18.80(62.00)
Meta-Llama-3.1-8B-Instruct	8B	✔	10.88	2.80	13.60	0.80	0.00	37.20(21.20)

Conclusion

- We propose **Knowledge-Orthogonal Reasoning (KOR),** aimed at minimizing reliance on domain-specific knowledge to evaluate reasoning abilities in out-of-distribution settings.

- We introduce **KOR-Bench**, a benchmark with five task categories: Operation, Logic, Cipher, Puzzle, and Counterfactual, **focusing on models' ability to apply new rule descriptions**.

- Results show that KOR-Bench demonstrates **strong differentiation and challenge**, advancing reasoning evaluation and supporting further research in AI reasoning and planning.

Project Page: https://kor-bench.github.io/