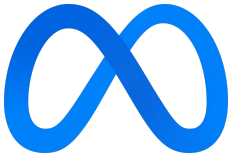


Law of the Weakest Link: Cross Capabilities of Large Language Models

Aston Zhang
Llama Team, Meta Generative AI

Joint work with M. Zhong, X. Wang, R. Hou, W. Xiong, C. Zhu, Z. Chen, L. Tan, C. Bi, M. Lewis, S. Popuri, S. Narang, M. Kambadur, D. Mahajan, S. Edunov, J. Han, and L. van der Maaten

Background: Development of LLMs

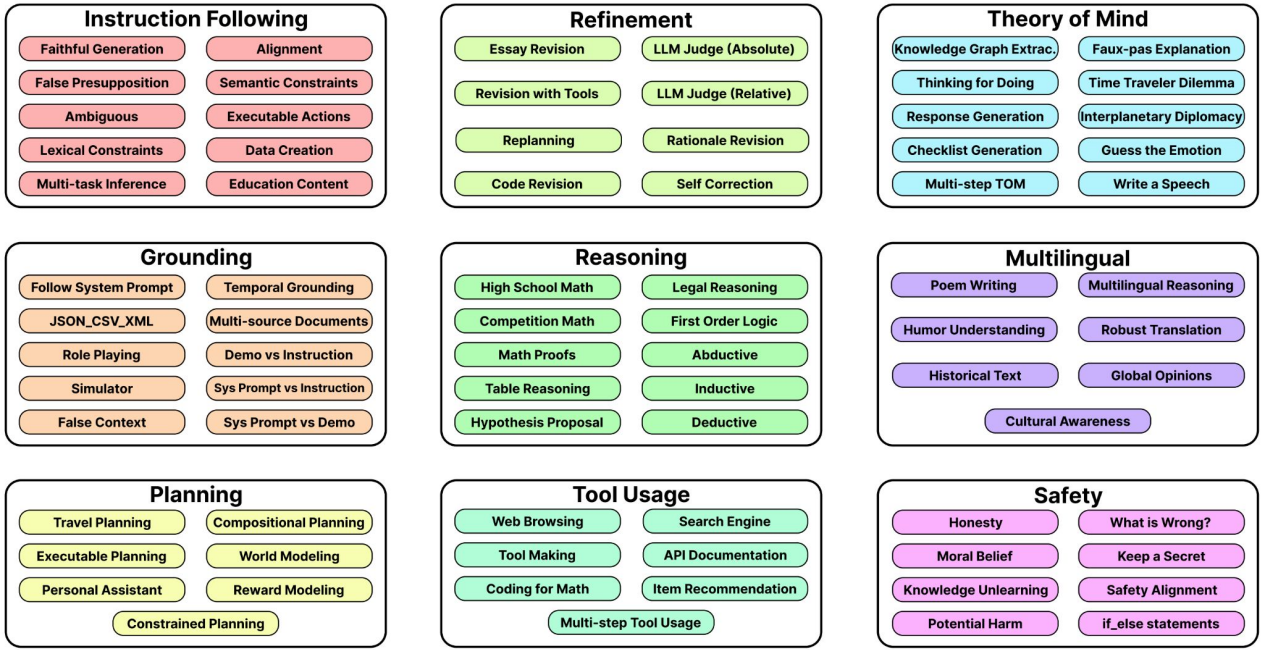
- Development of LLMs heavily focuses on individual capabilities
  - Data Mix + Supervised Fine-tuning
  - For example, data for post-training for Llama 3 is:

Dataset	% of examples	Avg. # turns	Avg. # tokens	Avg. # tokens in context	Avg. # tokens in final response
General English	52.66%	6.3	974.0	656.7	317.1
Code	14.89%	2.7	753.3	378.8	374.5
Multilingual	3.01%	2.7	520.5	230.8	289.7
Exam-like	8.14%	2.3	297.8	124.4	173.4
Reasoning and tools	21.19%	3.1	661.6	359.8	301.9
Long context	0.11%	6.7	38,135.6	37,395.2	740.5
Total	100%	4.7	846.1	535.7	310.4

Llama Team. The Llama 3 Herd of Models. 2024 Preprint.

Background: Evaluation of LLMs

- Evaluation of LLMs focuses on individual capabilities



Kim et al. The BIGGEN BENCH: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models. 2024 Preprint.

Motivation: Lack of Cross-Capability Exploration

- Many Real-World Tasks Require Cross Capabilities!!
  - **Examples**
    - Tool Use & Reasoning
      - Which direction has the total rainfall in Tokyo Japan been trending in the past 10 years? Explain it step by step
    - Long Context & Coding
      - Give me a basic understanding of what this web app does. I've provided all the HTML and some JS for the api data
  - **Definition**
    - We define these scenarios as **cross capabilities**
      - The intersection of multiple distinct capabilities across different types of expertise necessary to address complex, real-world tasks

Outline

- RQ Ⅰ: How can we comprehensively define individual and cross capabilities in LLMs?
- RQ II: How can we benchmark both individual and cross capabilities in LLMs?
- RQ III: What patterns exist in the relationship between individual and cross-capability performance in LLMs?
- RQ IV: How do performance shifts in individual capabilities impact cross-capability performance in LLMs?

RQ I

How can we comprehensively define individual and cross capabilities in LLMs?

Individual & Cross Capabilities

- 7 Single Capability
  - English
  - Coding
  - Reasoning
  - Image Recognition
  - Tool Use
  - Long Context
  - Multilingual (Spanish)
- 7 Cross Capability
  - Coding & Reasoning
  - Image Recognition & Reasoning
  - Long Context & Coding
  - Tool Use & Reasoning
  - Tool Use & Coding
  - Spanish & Reasoning
  - Spanish & Image Recognition

Taxonomy for Reasoning

- 7 Single Capabilities
  - English
  - Coding
  - **Reasoning**
  - Image Recognition
  - Tool Use
  - Long Context
  - Spanish

Examples for Reasoning

- Math Reasoning
  - Jane won the lottery and decided to spend some of the money. She spent $1.50 on the first day. She spent $3 on the second day. She spent $4.50 on the third day. She kept spending her winnings in the same pattern and then on the last day, she spent her remaining $300. How much did she win in the lottery?
- Social and Emotional Reasoning
  - Two chemists are sitting at a bar. The first chemist tells the bartender, “I’ll have some H2O.” The second chemist tells the bartender, “I will also have some water”. The first chemist tells the second chemist, “darn my murder plot failed”. Please explain this joke to me

Taxonomy for Image Recognition

- 7 Single Capabilities
  - English
  - Coding
  - Reasoning
  - **Image Recognition**
  - Tool Use
  - Long Context
  - Spanish

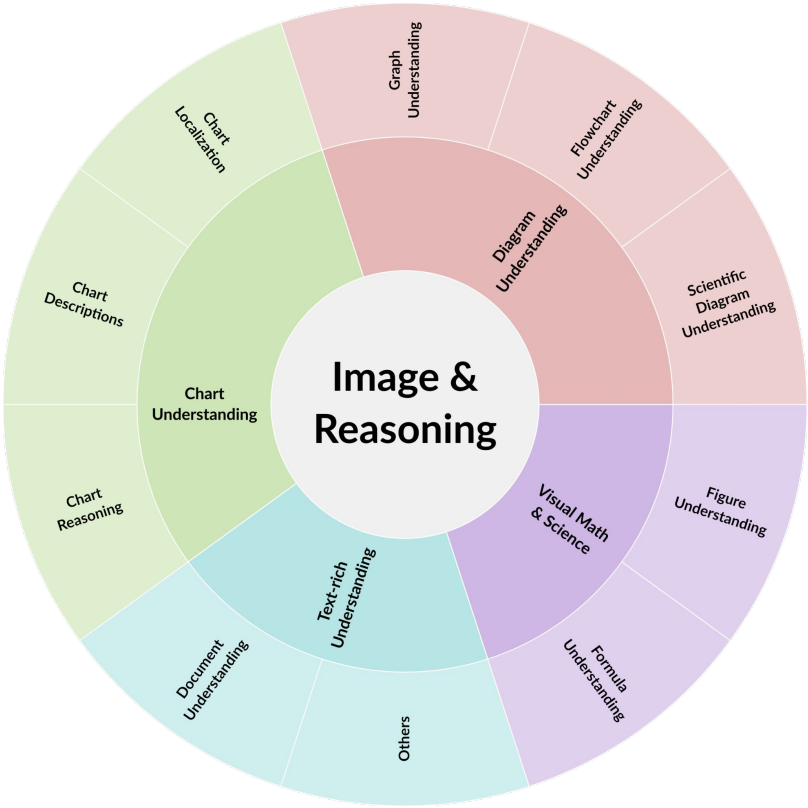
Examples for Image Recognition

- Object Recognition
  - How many of these dogs have floppy ears? How many of the floppy-eared dogs have black fur?

- Comprehensive Use Case
  - My bike is not ridable. What does it need to be fixed and how do I fix it?

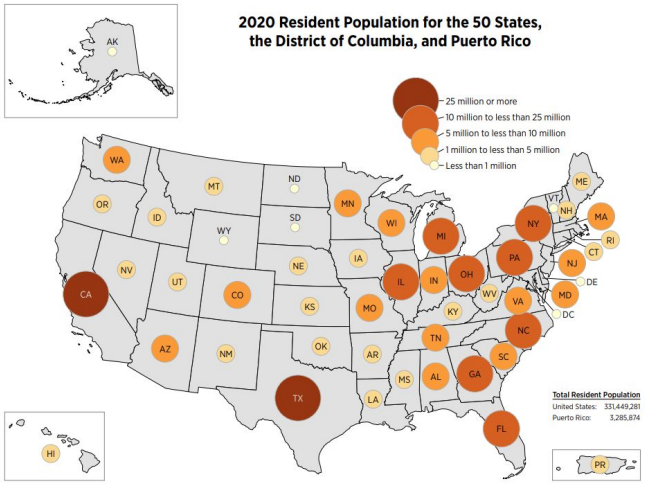
Taxonomy for Image Recognition & Reasoning

- 7 Cross Capabilities
  - Coding & Reasoning
  - **Image Recognition & Reasoning**
  - Long Context & Coding
  - Tool Use & Reasoning
  - Tool Use & Coding
  - Spanish & Reasoning
  - Spanish & Image



Examples for Image & Reasoning

- Chart Understanding
  - According to the chart, is the combined population of South Carolina and Alabama greater or less than the population of Texas?



- Text-rich Understanding
  - So what is the best strategy for the interview question in the image?

RQ II

How can we benchmark both individual and cross capabilities in LLMs?

CrossEval Benchmark

- Overview: what does our benchmark include?
  - Setting
    - Single-turn, open-ended, expert-annotated prompts
  - 1,400 prompts for 14 capabilities
  - Difficulty level
    - 10% Easy, 30% Medium, 60% Hard
  - Multiple reference examples for each prompt (used for evaluation)
    - 3 collected model responses (bad, medium, hard)
    - 2 reviews for each model response
      - Human rating + explanation
    - Overall, 4,200 model responses and 8,400 human ratings with explanations

CrossEval Benchmark

- Statistics for the prompts
  - 100 prompts for each capability
  - 76 Level-1 and 332 Level-2 categories

Capabilities		# Prompts	# L1 Categories	# L2 Categories
Individual	English	100	8	45
	Reasoning	100	8	36
	Coding	100	4	18
	Image Recognition	100	6	17
	Tool Use	100	5	23
	Long Context	100	3	14
	Spanish	100	8	45
Cross	Coding & Reasoning	100	4	19
	Image Recognition & Reasoning	100	4	10
	Tool Use & Coding	100	4	9
	Tool Use & Reasoning	100	3	30
	Long Context & Coding	100	5	13
	Spanish & Reasoning	100	8	36
	Spanish & Image Recognition	100	6	17

CrossEval Benchmark

- Why not annotate ground truth for each prompt?
  - Many open-ended queries do not have a single correct answer
    - How do you predict the 2024 U.S. election?
  - Many prompts remain challenging even for college-level expert annotators
    - Olympiad-level math or coding problem
  - For prompts related to tool use, the correct response can be dynamic
    - What is the temperature in Champaign today?
- Our solution
  - Reference examples specific to the current prompt for a more reliable LLM-as-a-Judge
    - Reference examples = 3 model responses + 6 human reviews

Reference Examples

Prompt

L1 Category: Mathematical Calculation
L2 Category: Discrete math & logic
Difficulty: Medium
Prompt: Jane won the lottery and decided to spend some of the money. She spent $1.50 on the first day. She spent $3 on the second day. She spent $4.50 on the third day. She kept spending her winnings in the same pattern and then on the last day, she spent her remaining $300. How much did she win in the lottery?

Model Response 1

Jane spent $1.50 on the first day, $3 on the second day, and $4.50 on the third day. So that is $1.50 + 3 + 4.50 = 9$ dollars in the first three days. She then spent $300 on the last day. So she spent $9 + 300 = 309$ dollars in total. The answer is 309.

Human Annotations

Rating 1: $1/5$
Explanation 1: The response does not understand that the user is asking for a pattern to be identified, and fails to understand that there were probably several days in between the third day and the final day. The response does not come up with a suitable formula and fails to correctly calculate the result; it instead just adds up the amounts specified in the prompt.

Rating 2: $2/5$
Explanation 2: The response was incorrect because it failed to identify the pattern in the given logic. The prompt indicates that Jane spent money according to a pattern observed in the first three days, but the response only calculated the total money spent on the first three days and the last day, ignoring the expenditures on the middle days.

Reference Examples

Model Response 3

Let's analyze the situation:
...

Now we can find the total amount she won:

$(1.5(199))(199+1)/2 + \$300 = \text{Total amount won}$
$(298.5)(200)/2 + \$300 = \text{Total amount won}$
$\$29850 + \$300 = \text{Total amount won}$
$\$30150 = \text{Total amount won}$

The final answer is $30150.

Human Annotations

Rating 1: $3/5$
Explanation 1: The response identifies and describes the correct pattern as well as providing the correct answer ($30150). But the explanation of the formula it provides is very confusing, particularly with the way that it separates the $300 on the last day. It is not certain that the user would be able to make use of the formula provided, but at least they would know what the pattern is, after reading this response.

Rating 2: $5/5$
Explanation 2: The response was correct because it identified the pattern of adding $1.5 on each subsequent day of spending. It correctly calculated the total amount spent before the last day using the appropriate formula and provided the accurate result of spending from the first day to the last day.

Inter-Rater Agreement

- Training Round on 30% Benchmark
  - Agreement Rate (1-5 scale)
    - 33.7% → 45.8% → 47.4%
  - K-Alpha
    - 0.48 → 0.66 → 0.73

- Inter-Rater Agreement on Full Benchmark after Training
  - Agreement Rate (1-5 scale)
    - 54.93%
  - K-Alpha
    - 0.76 (strong agreement)

Prompting LLMs to Evaluate Open-Ended Responses

- Previous Research Efforts
  - Issues with LLM-as-a-Judge
    - No reference, overly rely on the self-generated answer
    - No evaluation consistency
    - Tends to give higher scores, significantly above human ratings

- Our Methods
  - Multi-reference-based Prompting
    - 3 reference examples for each prompt
  - Point deduction-based Prompting
    - Summarize the typical point deductions from references
    - Apply them into model response evaluation

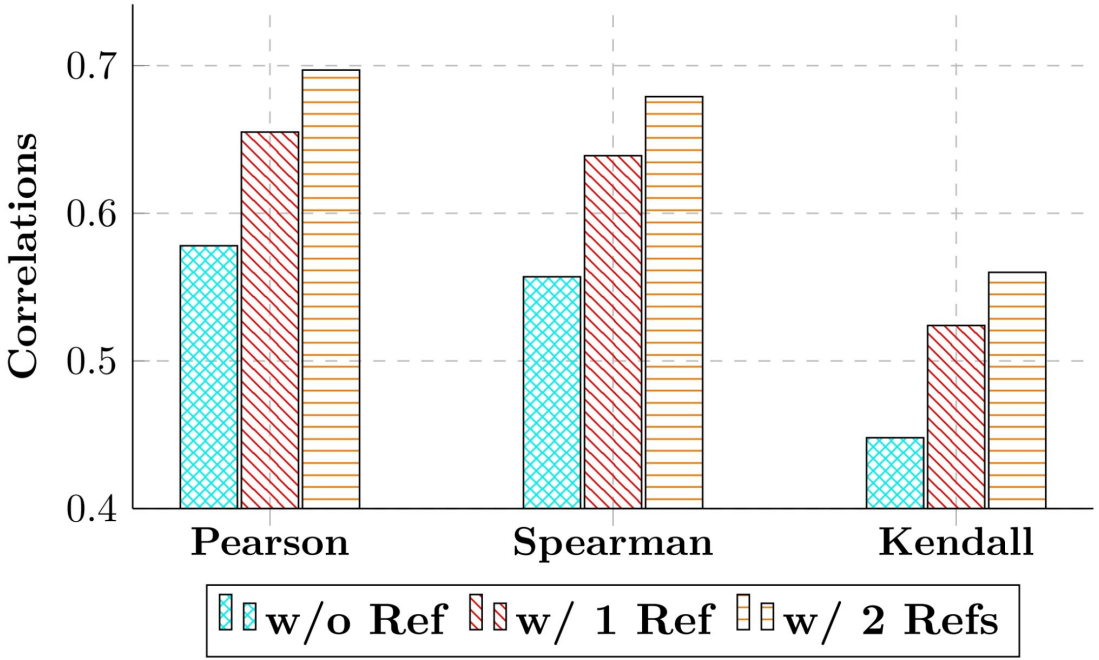
Correlations with Human Judgement

- Each LLM shows particular strengths in evaluating different capabilities
- Our methods can achieve nearly 0.7 Pearson correlations with expert judgements

Capabilities	GPT-4o mini	Llama 3.1 405B	Claude 3.5 Sonnet	GPT-4o-05-13
English	0.383	0.452	**0.516**	0.498
Reasoning	0.681	0.699	0.704	**0.731**
Coding	**0.627**	0.568	0.599	0.624
Image Recognition	0.576	–	0.733	**0.760**
Tool Use	0.587	0.609	**0.683**	0.629
Long Context	0.405	0.500	**0.609**	0.594
Spanish	0.552	0.536	**0.596**	0.594
Coding & Reasoning	0.618	0.600	0.623	**0.664**
Image Recognition & Reasoning	0.701	–	**0.819**	0.775
Tool Use & Coding	0.484	0.545	0.588	**0.639**
Tool Use & Reasoning	0.642	0.698	0.665	**0.729**
Long Context & Coding	0.524	0.535	**0.620**	0.593
Spanish & Reasoning	0.691	0.734	0.715	**0.772**
Spanish & Image Recognition	0.556	–	**0.752**	0.669
Overall Pearson (r)	0.621	–	0.696	**0.697**
Overall Spearman (r_s)	0.609	–	0.676	**0.679**
Overall Kendall (τ)	0.508	–	0.550	**0.560**

Ablation Study on the Reference Examples

- More Reference Examples → Better Correlations



RQ III

What patterns exist in the relationship between individual and cross-capability performance in LLMs?

Potential Patterns

Synergy Theory?

Cross performance > all individual

Potential Patterns

Synergy Theory?

Cross performance > all individual

Compensatory Mechanisms?

Stronger capabilities can compensate for weaker ones

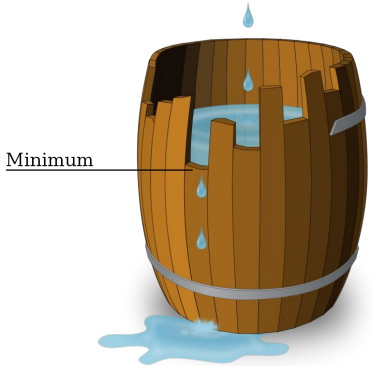
Potential Patterns

Synergy Theory?

Cross performance > all individual

Compensatory Mechanisms?

Stronger capabilities can compensate for weaker ones



Law of the Weakest Link?

Cross is limited by the weakest capability

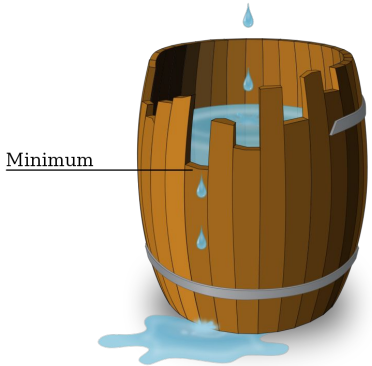
Potential Patterns

Synergy Theory?

Cross performance > all individual

Compensatory Mechanisms?

Stronger capabilities can compensate for weaker ones



Law of the Weakest Link?

Cross is limited by the weakest capability

Emergent Properties?

Cross performance is unpredictable

Results for Individual Capabilities

- CrossEval effectively differentiates advanced models
- Overall, Tool Use is the most challenging capability

	Individual Capabilities						
Models	**English**	**Reasoning**	**Coding**	**Image**	**Tool Use**	**Long Context**	**Spanish**
GPT-4o mini	73.64	69.31	71.17	65.23	–	76.18	74.51
GPT-4o	76.12	72.84	72.03	73.02	–	77.17	78.10
o1-mini	75.25	81.02	80.70	–	–	76.74	79.09
o1-preview	78.59	82.30	79.09	–	–	78.90	79.64
Claude 3 Haiku	63.87	56.81	61.64	51.00	–	69.68	67.95
Claude 3 Sonnet	69.19	62.88	66.09	56.56	–	72.40	69.43
Claude 3 Opus	68.94	66.22	69.68	61.76	–	74.69	74.01
Claude 3.5 Sonnet	75.00	71.54	**74.01**	68.57	–	74.32	76.12
Gemini 1.5 Flash	66.59	63.25	65.60	56.81	–	73.52	70.05
Gemini 1.5 Pro	71.91	70.61	69.56	69.56	–	**76.51**	74.26
Gemini 1.5 Pro Exp	**75.87**	**73.02**	69.56	**71.17**	–	75.37	**76.24**
Reka Edge	52.23	45.30	39.36	48.89	–	37.01	52.48
Reka Flash	63.87	62.63	57.68	56.38	–	55.82	68.07
Reka Core	71.54	68.69	62.38	56.94	–	60.90	73.77
Llama 3.1 8B	64.11	53.97	55.08	–	42.09	59.53	55.70
Llama 3.1 70B	68.82	62.88	65.47	–	47.04	68.82	64.48
Llama 3.1 405B	73.52	69.31	69.19	–	**47.90**	69.31	72.59

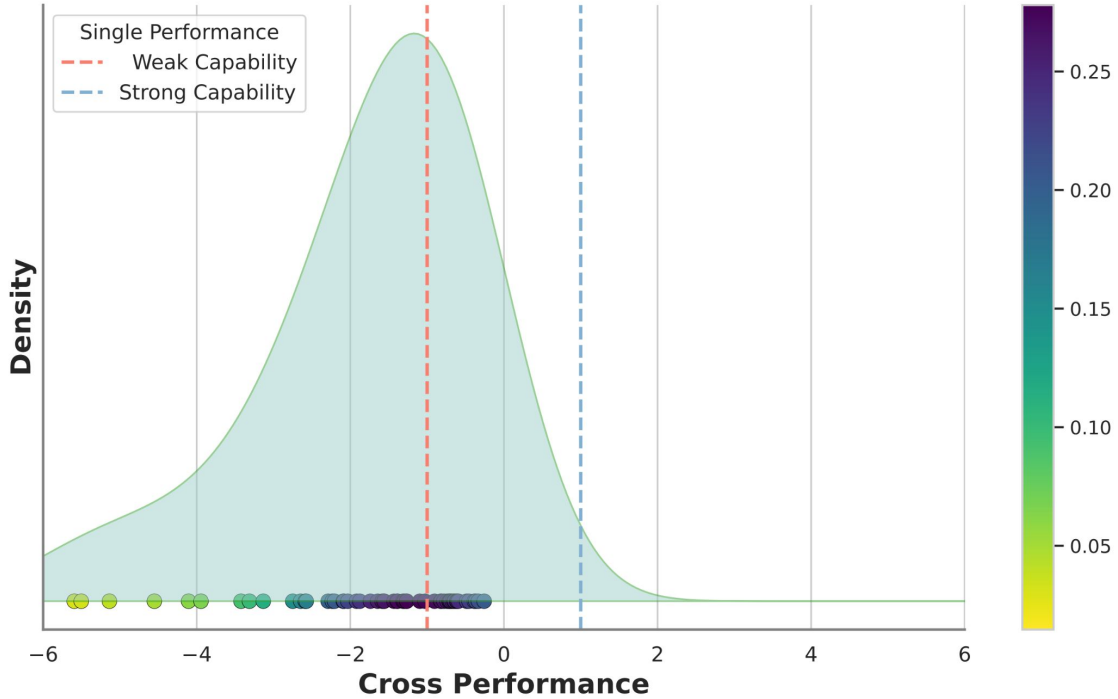
Results for Cross Capabilities

- Lower than the weaker capability: 38/58
- Between weak and strong, but closer to the weak: 20/58
- Closer to or better than strong: 0/58

	Cross Capabilities						
Models	**Coding & Rea.**	**Image & Rea.**	**Long & Coding**	**Spanish & Rea.**	**Spanish & Image**	**Tool & Coding**	**Tool & Rea.**
GPT-4o mini	72.03	65.60	65.10	69.56	65.10	–	–
GPT-4o	73.33	71.29	67.95	73.52	74.63	45.80	54.41
o1-mini	79.21	–	76.12	79.83	–	–	–
o1-preview	79.58	–	73.39	80.70	–	–	–
Claude 3 Haiku	58.05	49.88	58.67	57.80	52.85	–	–
Claude 3 Sonnet	61.14	54.71	58.79	60.77	60.52	–	–
Claude 3 Opus	63.37	53.84	58.17	67.33	64.11	–	–
Claude 3.5 Sonnet	**71.41**	**69.43**	65.72	70.55	69.81	–	–
Gemini 1.5 Flash	64.73	51.74	62.13	65.10	53.10	–	–
Gemini 1.5 Pro	69.68	67.95	**65.97**	69.56	62.26	–	–
Gemini 1.5 Pro Exp	67.33	69.06	**65.97**	**71.54**	**70.18**	–	–
Reka Edge	41.34	28.60	20.43	40.97	45.06	–	–
Reka Flash	56.94	43.45	37.63	59.66	55.82	–	–
Reka Core	63.62	46.66	41.25	68.01	54.71	–	–
Llama 3.1 8B	55.08	–	45.06	46.42	–	46.91	43.82
Llama 3.1 70B	67.21	–	50.50	59.41	–	50.25	49.45
Llama 3.1 405B	66.96	–	54.58	64.48	–	**52.23**	**51.74**

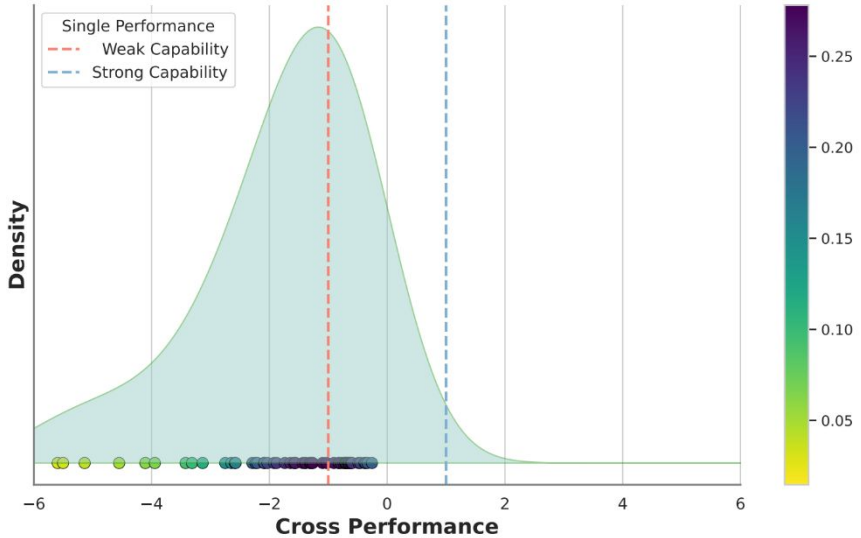
“Law of the Weakest Link” in LLMs

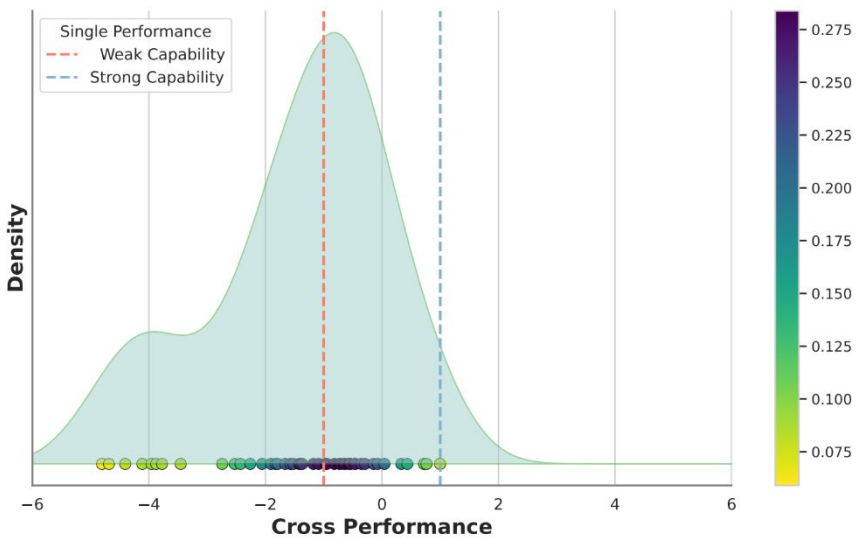
- Remap **weak** to -1, **strong** to 1, plot the distribution of cross performance
- Cross performance is significantly limited by (clusters around) the weak capability



“Law of the Weakest Link” is Evaluator-Agnostic

- Left: GPT-as-a-Judge
- Right: Claude-as-a-Judge





RQ IV

How do performance shifts in individual capabilities impact cross performance?

How to Enhance Specified Individual Capability?

- Principle-based System Prompting
  - Learn from the error cases from the specified capability for the given model
  - Iteratively update the principles
  - Finally, 10 principles for each capability
  - Use the generated principles as the system prompt to instruct the model

For each iteration, choose **ONE of the following actions:**

1. **ADD**
   - Introduce a new principle that isn't currently listed.
2. **REPLACE**
   - Replace a less significant principle with a new one.
   - Clearly specify which principle is being replaced.
3. **REVISE**
   - Enhance the principles by making them more detailed and specific.
4. **KEEP**
   - If the current instance is already covered by existing principles, leave the guideline unchanged.

Case Study for the Generated Principles

Principle 7: For Scientific Reasoning and Empirical Analysis
- **Verify the Existence of Citations**:
  1. Confirm all citations are based on actual research papers, cross-referencing with recognized academic databases.
  2. Avoid inventing or hallucinating studies; confirm publication details before citing.
- **Summarize Study Findings Accurately**:
  1. Provide specific results and data points from studies to back claims.
  2. Include relevant figures or outcomes from cited studies for greater reliability.
- **Incorporate Empirical Evidence**:
  1. Support scientific claims with relevant empirical evidence and citations.
  2. Avoid overgeneralizations; use specific examples or case studies.

“Law of the Weakest Link” Persists after Enhancement

- Take Image (**weak**) & Reasoning (**strong**) in Claude 3 as an example
  - Weak ↓ 0.99 + Strong ↑ 2.85 → Cross ↓ 3.46
  - Weak ↑ 3.71 + Strong ↓ 1.36 → Cross ↑ 4.58
- 90% changes in cross performance closely follow the trends of the weaker capability

Models	*Individual Capabilities*			*Cross Capabilities*		
	Reasoning	Image Recognition	Spanish	Image & Rea.	Spanish & Rea.	Spanish & Image
Claude 3 Haiku	56.81	51.00	67.95	49.88	57.80	52.85
+ Reasoning	**59.66**	50.01	**68.20**	46.42	**59.04**	52.11
+ Image	55.45	**54.71**	64.98	**54.46**	57.55	**55.08**
+ Spanish	55.20	53.59	67.21	50.13	56.81	53.72
Gemini 1.5 Flash	63.25	56.81	70.05	51.74	65.10	53.10
+ Reasoning	**66.71**	62.50	**71.29**	**54.46**	**66.59**	59.04
+ Image	59.91	**63.00**	69.43	51.61	62.13	**61.76**
+ Spanish	61.39	61.89	69.06	52.60	64.86	58.42

Takeaways

- Cross capability is the critical oversight of the development and evaluation of LLMs
- New testbed for evaluating cross capability
  - New benchmark + new LLM-as-a-Judge framework
- Identify “Law of the Weakest Link” in LLMs
  - For both static and dynamic evaluation
- Implications
  - Capture and enhance the weakest capabilities should be a priority for future development
  - Standalone LLMs are insufficient for real-world tasks due to inevitable weak points

Thank You!