

CS11-711 Advanced NLP

Evaluation and Benchmarks

Carnegie
Mellon
University



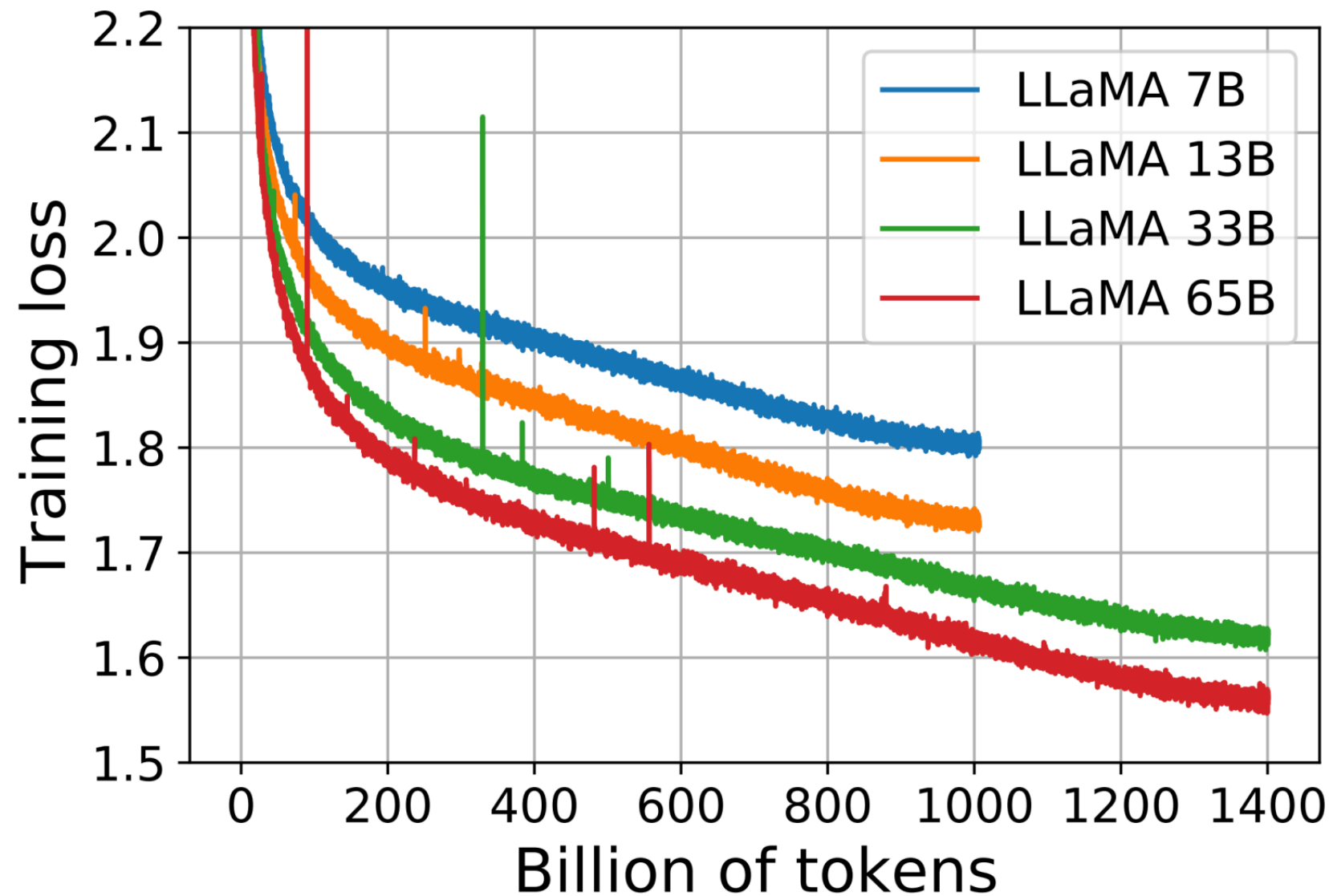
Recap

- Learning: pre-training, fine-tuning
- Inference: decoding algorithms
- Today: How do we evaluate how good our system is?
 - How do we choose between two systems?

Recap: Loss-Based Evaluation

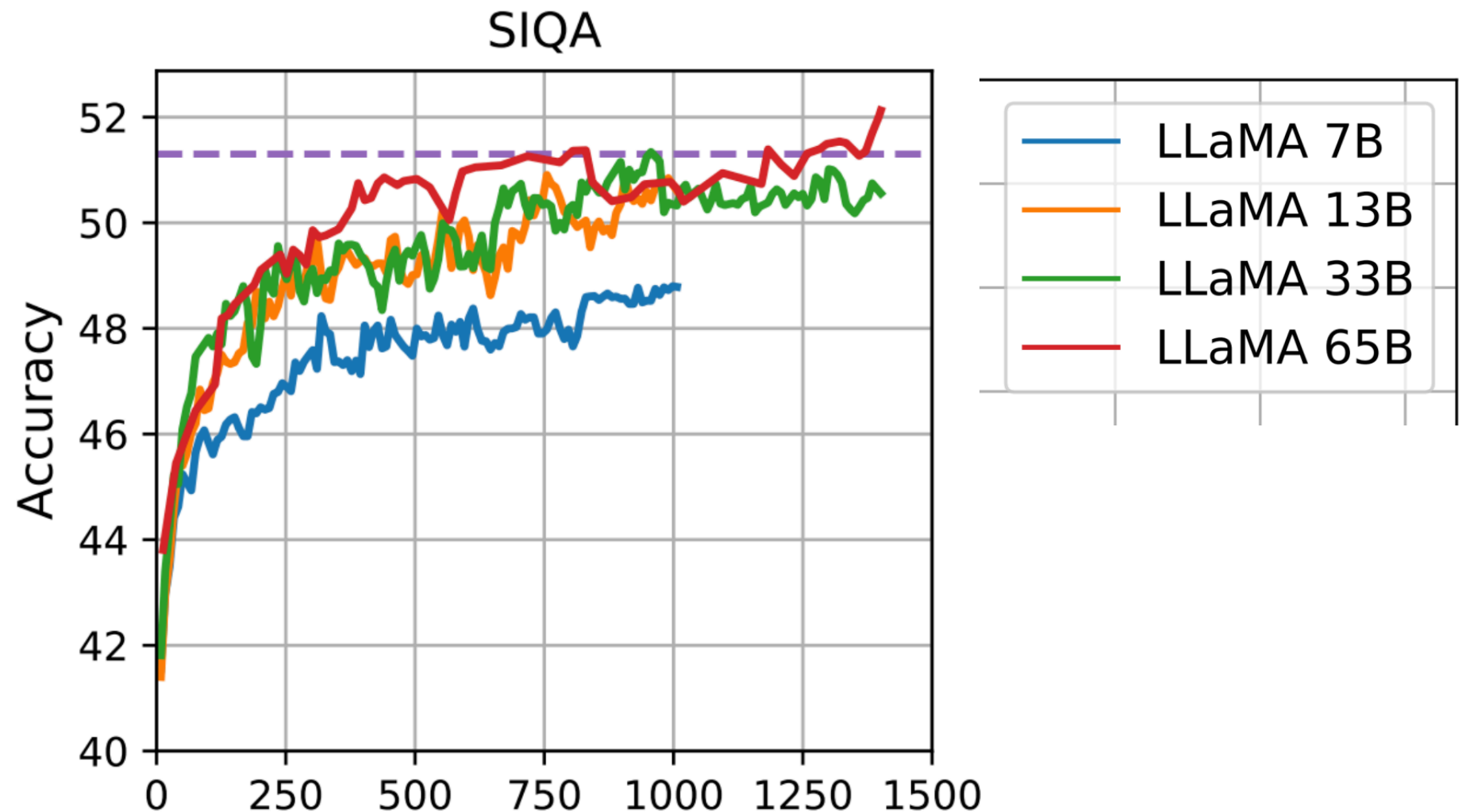
- Let p_θ be a language model
- Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a dataset
- Let $\mathcal{L}(\theta; \mathcal{D})$ be the loss function
 - Typically $-\log p_\theta(y | x)$
- Then we can use the training loss, validation loss, or test loss to evaluate the model

Recap: Loss-Based Evaluation



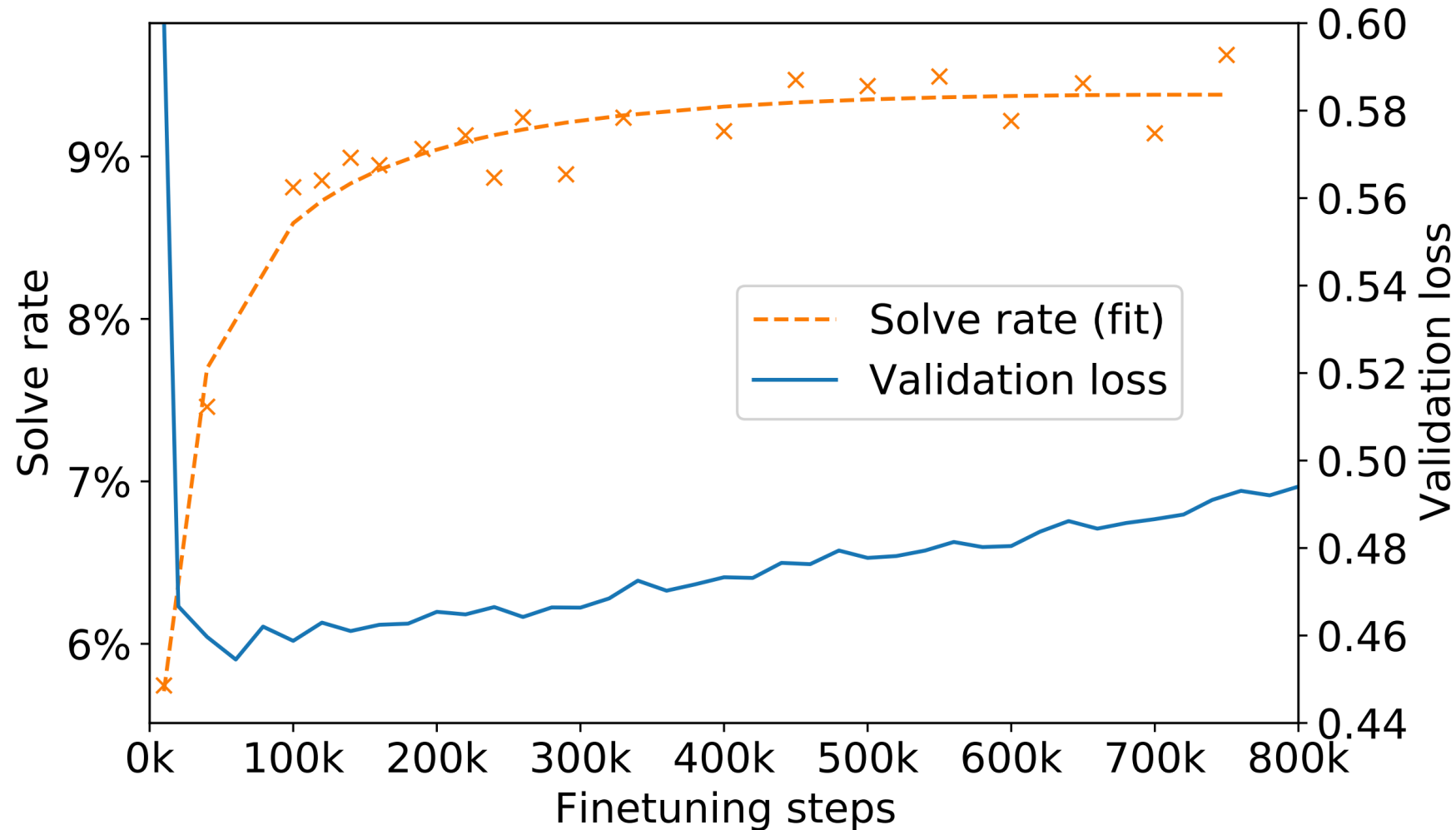
Claim: “Llama 33B is better than llama 13B”

Loss isn't always enough



Claim: “Llama 33B is better than llama 13B” (?)

Loss isn't always enough



Li et al. 2022 (DeepMind): *Competition-Level Code Generation with AlphaCode*

Loss isn't always enough

- Loss measures how well the model fits the data distribution
- It does not directly measure how well a model performs tasks

Task metrics and downstream evaluation

- Let $g(p_\theta, x)$ denote running an inference algorithm
- Let $m(y, \hat{y}) \rightarrow \mathbb{R}$ be a task metric (e.g., solve rate)
- Let $p(D)$ be a distribution over task datasets
- We want the model to have high expected task performance

$$\mathbb{E}_{D \sim p(D)} \mathbb{E}_{x, y \sim D} \mathbb{E}_{\hat{y} \sim g(p_\theta, x)} [m(y, \hat{y})]$$

- Each dataset D represents a “downstream task”, i.e. a task that we use the model for downstream of its training

How do we evaluate?

- In practice, we use a finite number of *evaluation benchmarks*. Each benchmark typically consists of:
 - A dataset D (sometimes only inputs x)
 - A task metric m (typically $m(y, \hat{y})$)
- We want the benchmarks to cover the breadth and depth of functionality that we want in our system.
 - For a machine translation system, we want various machine translation benchmarks
 - For a language model API, we want benchmarks related to the many capabilities that customers want from the API

This lecture

- Intro/evaluation setup
- Properties of good benchmarks
 - Example benchmarks
 - Example task metrics
- Can I trust the benchmark score?

Example: Llama 3 paper

Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 ₍₀₁₂₅₎	GPT-4o	Claude 3.5 Sonnet
General	MMLU _(5-shot)	69.4	72.3	61.1	83.6	76.9	70.7	87.3	82.6	85.1	89.1	89.9
	MMLU _(0-shot, CoT)	73.0	72.3 [△]	60.5	86.0	79.9	69.8	88.6	78.7 [△]	85.4	88.7	88.3
	MMLU-Pro _(5-shot, CoT)	48.3	—	36.9	66.4	56.3	49.2	73.3	62.7	64.8	74.0	77.0
	IFEval	80.4	73.6	57.6	87.5	72.7	69.9	88.6	85.1	84.3	85.6	88.0
Code	HumanEval _(0-shot)	72.6	54.3	40.2	80.5	75.6	68.0	89.0	73.2	86.6	90.2	92.0
	MBPP EvalPlus _(0-shot)	72.8	71.7	49.5	86.0	78.6	82.0	88.6	72.8	83.6	87.8	90.5
Math	GSM8K _(8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6	96.8	92.3 [◇]	94.2	96.1	96.4 [◇]
	MATH _(0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1	73.8	41.1	64.5	76.6	71.1
Reasoning	ARC Challenge _(0-shot)	83.4	87.6	74.2	94.8	88.7	83.7	96.9	94.6	96.4	96.7	96.7
	GPQA _(0-shot, CoT)	32.8	—	28.8	46.7	33.3	30.8	51.1	—	41.4	53.6	59.4
Tool use	BFCL	76.1	—	60.4	84.8	—	85.9	88.5	86.5	88.3	80.5	90.2
	Nexus	38.5	30.0	24.7	56.7	48.5	37.2	58.7	—	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	—	—	90.5	—	—	95.2	—	95.2	90.5	90.5
	InfiniteBench/En.MC	65.1	—	—	78.2	—	—	83.4	—	72.1	82.5	—
	NIH/Multi-needle	98.8	—	—	97.5	—	—	98.1	—	100.0	100.0	90.8
Multilingual	MGSM _(0-shot, CoT)	68.9	53.2	29.9	86.9	71.1	51.4	91.6	—	85.9	90.5	91.6

Table 2 Performance of finetuned Llama 3 models on key benchmark evaluations. The table compares the performance of

- “Key benchmarks” in the Llama 3 paper

Example: Gemini 3.1 (Feb 2026)

Benchmark		Gemini 3.1 Pro Thinking (High)	Gemini 3 Pro Thinking (High)	Sonnet 4.6 Thinking (Max)	Opus 4.6 Thinking (Max)	GPT-5.2 Thinking (xhigh)	GPT-5.3-Codex Thinking (xhigh)
Humanity's Last Exam Academic reasoning (full set, text + MM)	No tools	44.4%	37.5%	33.2%	40.0%	34.5%	—
	Search (blocklist) + Code	51.4%	45.8%	49.0%	53.1%	45.5%	—
ARC-AGI-2 Abstract reasoning puzzles	ARC Prize Verified	77.1%	31.1%	58.3%	68.8%	52.9%	—
GPQA Diamond Scientific knowledge	No tools	94.3%	91.9%	89.9%	91.3%	92.4%	—
Terminal-Bench 2.0 Agentic terminal coding	Terminus-2 harness Other best self-reported harness	68.5% —	56.9% —	59.1% —	65.4% —	54.0% 62.2% (Codex)	64.7% 77.3% (Codex)
SWE-Bench Verified Agentic coding	Single attempt	80.6%	76.2%	79.6%	80.8%	80.0%	—
SWE-Bench Pro (Public) Diverse agentic coding tasks	Single attempt	54.2%	43.3%	—	—	55.6%	56.8%
LiveCodeBench Pro Competitive coding problems from Codeforces, ICPC, and IOI	Elo	2887	2439	—	—	2393	—
SciCode Scientific research coding		59%	56%	47%	52%	52%	—
APEX-Agents Long horizon professional tasks		33.5%	18.4%	—	29.8%	23.0%	—
GDPval-AA Elo Expert tasks		1317	1195	1633	1606	1462	—
τ2-bench Agentic tool use	Retail	90.8%	85.3%	91.7%	91.9%	82.0%	—
	Telecom	99.3%	98.0%	97.9%	99.3%	98.7%	—
MCP Atlas Multi-step workflows using MCP		69.2%	54.1%	61.3%	59.5%	60.6%	—
BrowseComp Agentic search	Search + Python + Browse	85.9%	59.2%	74.7%	84.0%	65.8%	—
MMMU Pro Multimodal understanding and reasoning	No tools	80.5%	81.0%	74.5%	73.9%	79.5%	—
MMMLU Multilingual Q&A		92.6%	91.8%	89.3%	91.1%	89.6%	—
MRCR v2 (8-needle) Long context performance	128k (average)	84.9%	77.0%	84.9%	84.0%	83.8%	—
	1M (pointwise)	26.3%	26.3%	Not supported	Not supported	Not supported	—

Example: MMLU

- Multiple-choice questions in 57 subjects
 - Model generates A, B, C, or D
 - Metric: exact match; is the answer the same as the one in the dataset
 - $m(y, \hat{y}) = 1[y = \hat{y}]$

Microeconomics	One of the reasons that the government discourages and regulates monopolies is that	
	(A) producer surplus is lost and consumer surplus is gained.	✗
	(B) monopoly prices ensure productive efficiency but cost society allocative efficiency.	✗
	(C) monopoly firms do not engage in significant research and development.	✗
	(D) consumer surplus is lost with higher prices and lower levels of output.	✓

Figure 3: Examples from the Microeconomics task.

Conceptual Physics	When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is	
	(A) 9.8 m/s^2	✓
	(B) more than 9.8 m/s^2	✗
	(C) less than 9.8 m/s^2	✗
	(D) Cannot say unless the speed of throw is given.	✗

Example: HumanEval

- HumanEval: LeetCode-style Python problems
 - Model generates code
 - Metric: execute the code and check whether it passes all test cases

$$m(y, \hat{y}) = \prod_{j=1}^{numtests} 1[\hat{y} \text{ passes test } j]$$

$$\text{pass @ } K(y, \{y_1, \dots, y_K\}) = \max(m(y, y_1), \dots, m(y, y_K))$$

```
def solution(lst):  
    """Given a non-empty list of integers, return the sum of all of the odd elements  
    that are in even positions.  
  
    Examples  
    solution([5, 8, 7, 1]) ==>12  
    solution([3, 3, 3, 3, 3]) ==>9  
    solution([30, 13, 24, 321]) ==>0  
    """  
    return sum(lst[i] for i in range(0, len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

Example: GSM8k

- Grade school mathematics questions
- Model generates a chain-of-thought and then an answer
- Metric: check whether the answer is the same as the one in the dataset

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?

Solution: Tina buys 3 12-packs of soda, for $3 \times 12 = 36$ sodas

6 people attend the party, so half of them is $6/2 = 3$ people

Each of those people drinks 3 sodas, so they drink $3 \times 3 = 9$ sodas

Two people drink 4 sodas, which means they drink $2 \times 4 = 8$ sodas

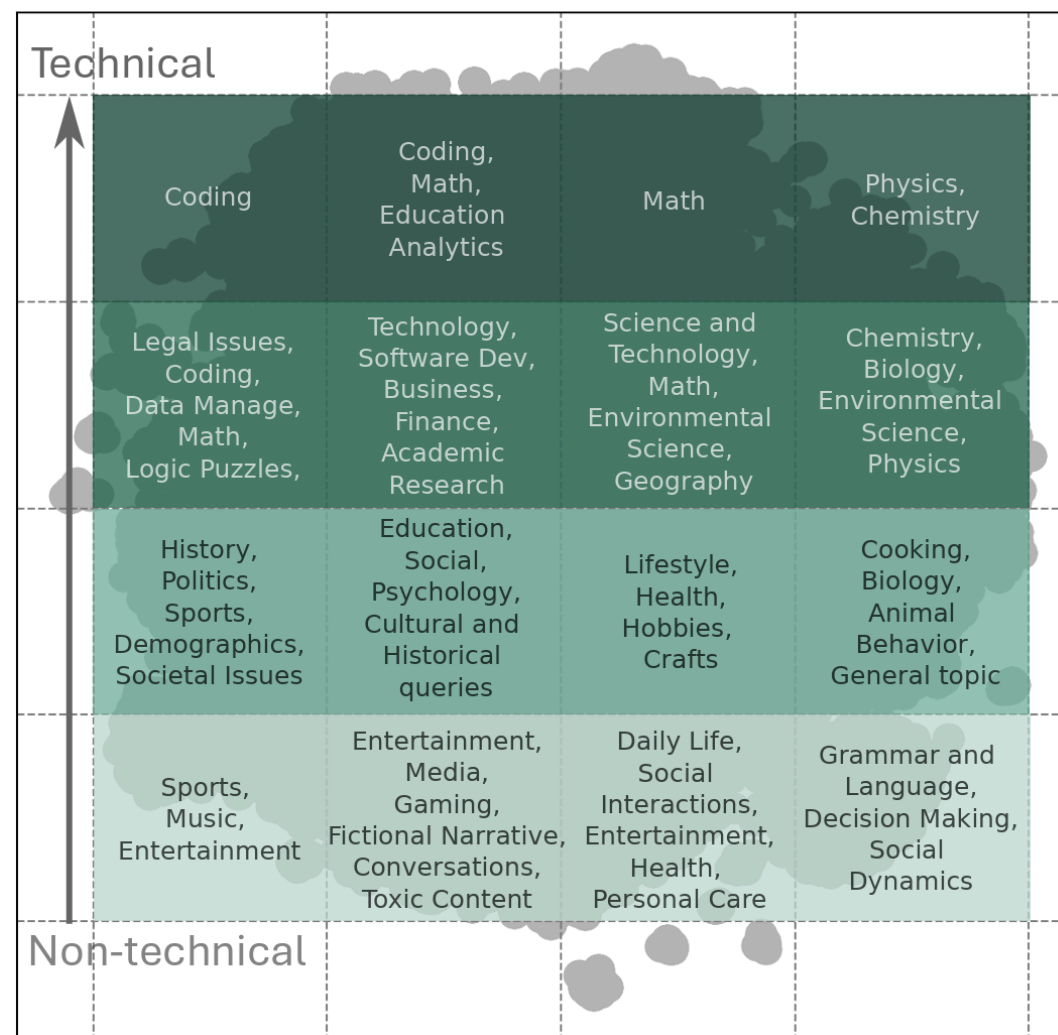
With one person drinking 5, that brings the total drank to $5 + 9 + 8 + 3 = 25$ sodas

As Tina started off with 36 sodas, that means there are $36 - 25 = 11$ sodas left

Final Answer: 11

Discussion: properties of good benchmarks?

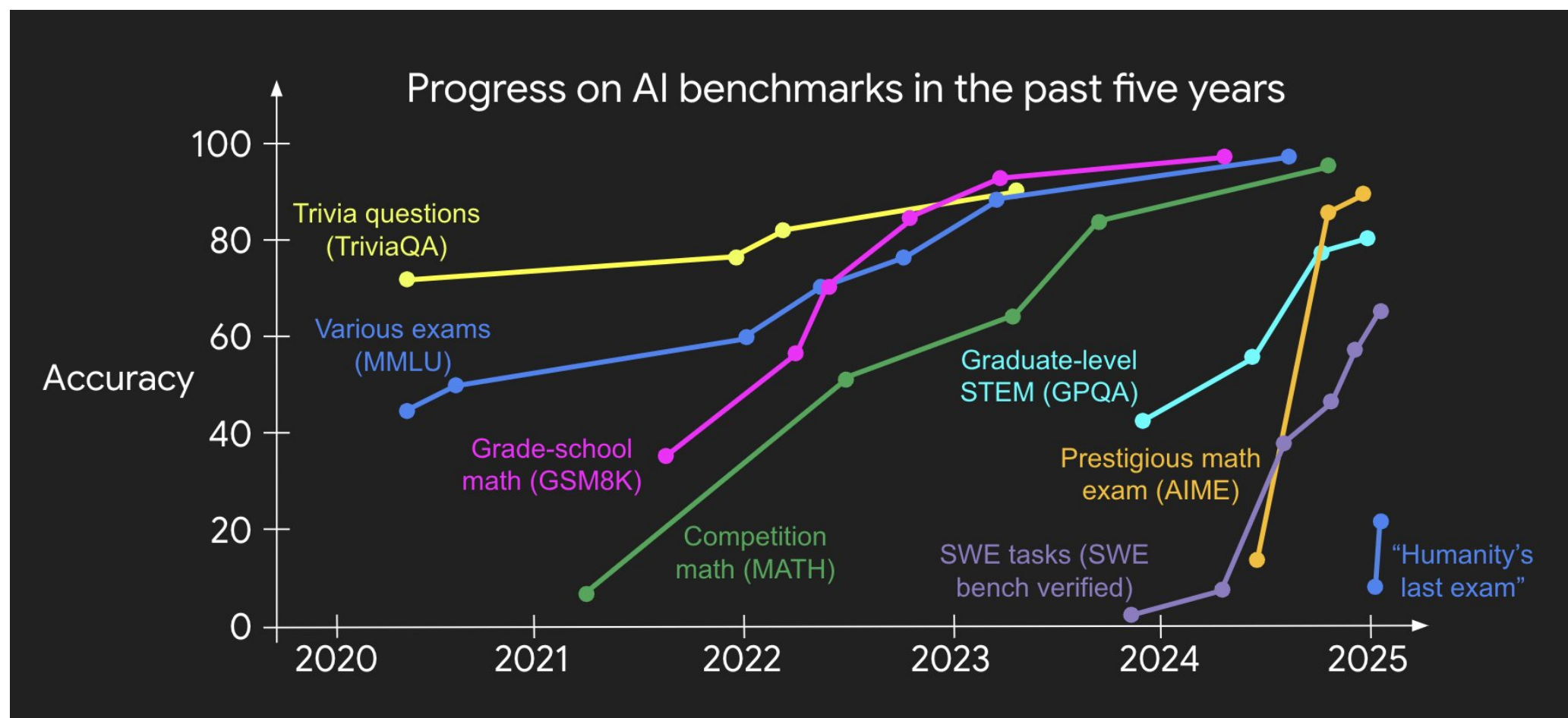
- Breadth/diversity
- Evaluates all desired input-output functionality



Prompt types from active users [Ni et al 2024]

Discussion: properties of good benchmarks?

- Depth/difficulty
- Examples are difficult enough to distinguish good and bad models.



Discussion: properties of good benchmarks?

- Utility
- Task is a proxy for desired functionality

MATH Dataset (Ours)

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

Solution: Complete the square by adding 1 to each side. Then $(x + 1)^2 = 1 + i = e^{\frac{i\pi}{4}} \sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}} \sqrt[4]{2}$. The desired product is then $(-1 + \cos(\frac{\pi}{8}) \sqrt[4]{2})(-1 - \cos(\frac{\pi}{8}) \sqrt[4]{2}) = 1 - \cos^2(\frac{\pi}{8}) \sqrt{2} = 1 - \frac{(1 + \cos(\frac{\pi}{4}))}{2} \sqrt{2} = \boxed{\frac{1 - \sqrt{2}}{2}}$.

Why do we need an LLM
that is good at solving
math word problems?

Foundation for **more complex tasks**
(e.g., financial analysts)

Useful for **real users with similar questions**
(e.g., tutoring high school math)

Medium for **addressing research questions**
(e.g., Can an AI model reason?)

Discussion: properties of good benchmarks?

- Robustness

Question Context: James *leaves* home for shopping. He walks 4 miles/hour in the first 2 hours.

Question 1

Then he increased his speed to 5 miles/hour. After one more hour, how far is he from home?



Answer:

Step 1: In the first 2 hours, James walks $2 * 4 = 8$ miles. ✓

Step 2: For the remaining one hour, he walks $1 * 5 = 5$ miles. ✓

Step 3: Therefore, James is $8 + 5 = 13$ miles away from home. ✓

Question 2

Then he realized that he forgot something at home and had to *return* and increased his speed to 5 miles/hour. After one more hour, how far is he from home?



Answer:

Step 1: In the first 2 hours, he walks $2 * 4 = 8$ miles. ✓

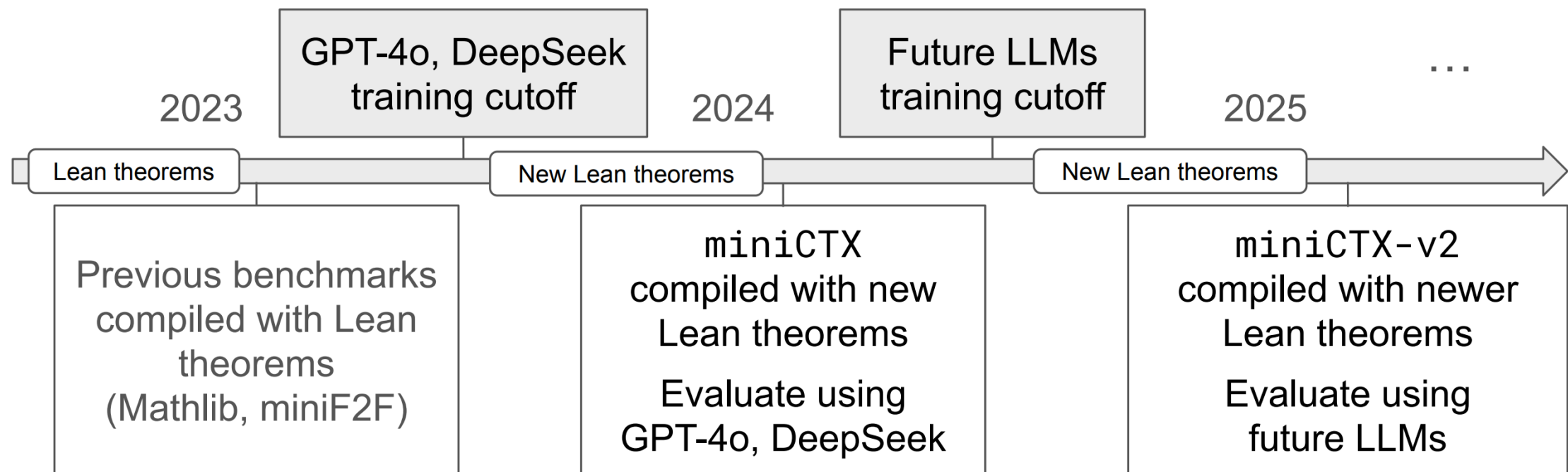
Step 2: In the third hour, he walks $1 * 5 = 5$ miles. ✓

Step 3: Therefore, James is $8 + 5 = 13$ miles away from home. ✗ (It should be $8 - 5 = 3$ miles due to "return" yielding the opposite directions.)

GSM-Plus: Robustness to input variations

Discussion: properties of good benchmarks?

- Data contamination



“Temporal split” in miniCTX [Hu et al]

Discussion: properties of good benchmarks?

- Efficient to evaluate
 - Hire an expert human to read a long chat and decide if it's correct
 - Expensive, slow, error prone
- ...
- Check whether a generated integer answer matches the answer in the dataset
 - Cheap, fast, exact, but simplifies the task

Automatically evaluating complex text

Question: What are some business etiquette norms when doing business in Japan?

Assistant A:

Here are some business etiquette norms to keep in mind when doing business in Japan:

1. Bowing: Bowing is a common greeting in Japan. It is a sign of respect and should be reciprocated.
2. Business cards: Business cards are exchanged at the beginning of a meeting. It is important to receive and give business cards with both hands and to take a moment to read the card before putting it away.
3. Punctuality: Being on time is important in Japan. Arrive a few minutes early to show respect for your colleagues.
4. Dress code: Dress conservatively and professionally. Men should wear dark suits and ties, while women should wear conservative business attire.
5. Language: Learn a few basic Japanese phrases to show respect for the culture. It is also important to speak clearly and avoid using slang or idioms.
6. Gift giving: Gift giving is a common practice in Japan. Bring a small gift, such as a box of chocolates or a small souvenir from your home country.
7. Dining etiquette: When dining with colleagues, wait for the host to begin eating before you start. Use chopsticks correctly and avoid leaving food on your plate.

Remember, showing respect for Japanese culture and customs is key to building strong business relationships in Japan.

Assistant B:

When doing business in Japan, it is important to be aware of and respect the country's business etiquette norms. Some of these norms include:

1. Greetings: A handshake and bow is the standard greeting in Japan. It is important to bow deeply and to offer a firm handshake.
2. Business cards: In Japan, business cards are exchanged after the initial greeting. It is important to handle the card with respect and to read it carefully before placing it on the table.
3. Meetings: Meetings in Japan are typically formal and conducted in a hierarchical manner. It is important to address people by their titles and to avoid direct eye contact.
4. Punctuality: Punctuality is highly valued in Japan. It is important to arrive on time for meetings and to be prepared in advance.
5. Gift giving: Gift giving is a common practice in Japan, but it is important to be mindful of the occasion and the relationship. Avoid giving gifts that are considered too personal or extravagant.
6. Dining: Dining is an important part of doing business in Japan. It is important to be aware of the proper table manners and to avoid slurping or making loud noises while eating.
7. Communication: Communication in Japan is typically indirect and reserved. It is important to listen carefully and to avoid confrontational language.

Overall, it is important to be respectful, punctual, and mindful of hierarchy when doing business in Japan.

Automatically evaluating complex text

- Option 1: word overlap
- **Rouge**: Count the word overlap between the prediction & ref. answer

Question:

What are the benefits of regular exercise?

Prediction (5 words):

Exercise improves mood and health.

Reference Answer (6 words):

Regular exercise benefits health and mood.

- **Rouge-1**: “exercise”, “health”, “and”, “mood” overlaps -> Precision: 4/5, Recall: 4/6, F1: 0.727
- **Rouge-2**: No overlap! -> Precision: 0/5, Recall: 0/6, F1: 0.0 **Is this desirable?**
- **Rouge-L**: [“exercise”, “mood”] and [“exercise”, “and”] are common subsequences -> Precision: 2/5, Recall: 2/6, F1: 0.364

Discuss: Pros? Cons?

Automatically evaluating complex text

- Option 2: use a LLM (“LLM as a judge”)
 - Example: $z \sim LLM(x, y_A, y_B; \text{prompt})$
 - Prompt: “Which output is better, A or B? Think carefully.”
 - Then we can compute the *win-rate* of model A versus model B.

Discuss: Pros? Cons?

Automatically evaluating complex text

- Option 2: use a LLM (“LLM as a judge”)

GPT-4 Judgment (when A is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant A's answer is more detailed and organized, covering a wider range of topics such as dress code and language, which are not mentioned in Assistant B's response. Assistant A also provides more specific advice on how to handle certain situations, such as exchanging business cards and dining etiquette. Therefore, Assistant A's answer is more comprehensive and useful for someone looking to understand Japanese business etiquette.

GPT-4 Judgment (when B is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant B's answer is more detailed and covers a wider range of topics, such as addressing people by their titles, avoiding direct eye contact, and being mindful of communication styles. Assistant A's answer is more concise but lacks some of the depth found in Assistant B's response. Therefore, Assistant B's answer is better in this case.

Discuss: Pros? Cons?

High agreement with humans [Zheng et al 2023]

Table 5: Agreement between two types of judges on MT-bench. “G4-Pair” and “G4-Single” denote GPT-4 with pairwise comparison and single-answer grading respectively. The single-answer grading can be converted into pairwise comparison results for calculating the agreement. We report two setups: “S1” includes non-tie, tie, and inconsistent (due to position bias) votes and counts inconsistent as tie; “S2” only includes non-tie votes. The agreement between two random judges under each setup is denoted as “R=”. The top value in each cell is the agreement, and the bottom gray value is #votes.

Setup	S1 (R = 33%)		S2 (R = 50%)	
Judge	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1138	66% 1343	97% 662	85% 859
G4-Single	-	60% 1280	-	85% 739
Human	-	63% 721	-	81% 479

(a) First Turn

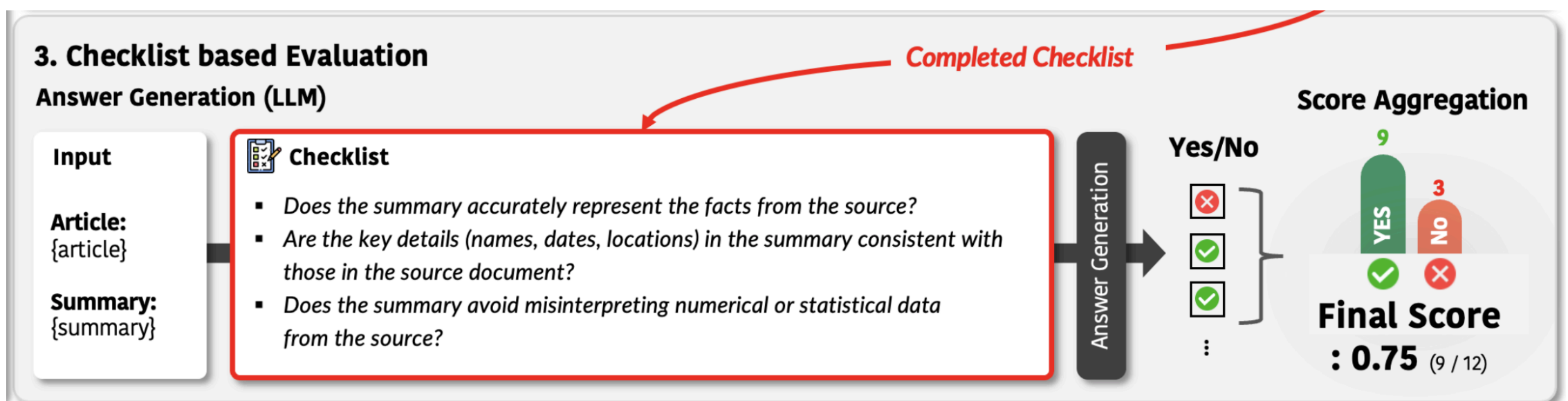
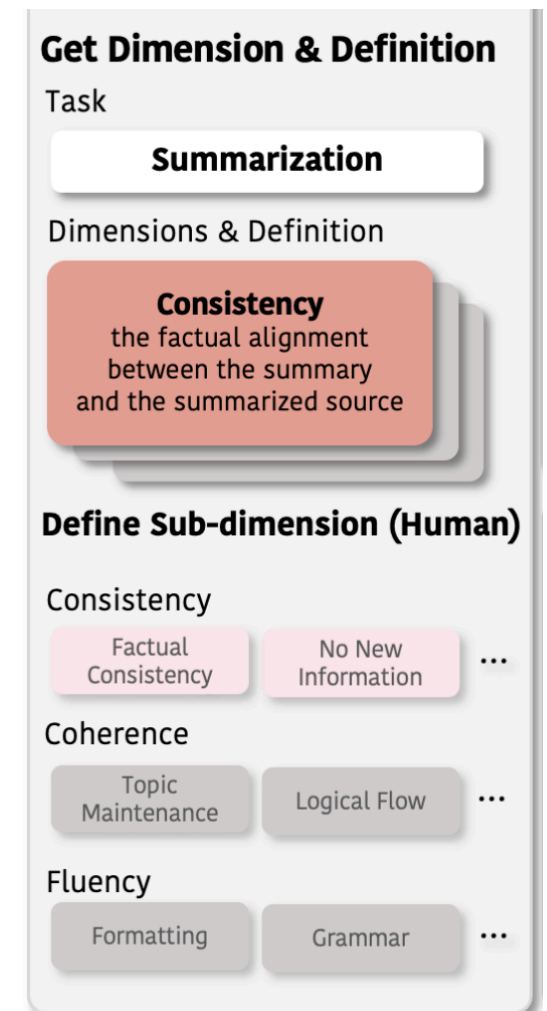
Setup	S1 (R = 33%)		S2 (R = 50%)	
Judge	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1161	66% 1325	95% 727	85% 864
G4-Single	-	59% 1285	-	84% 776
Human	-	67% 707	-	82% 474

(b) Second Turn

Checklist-based evaluation

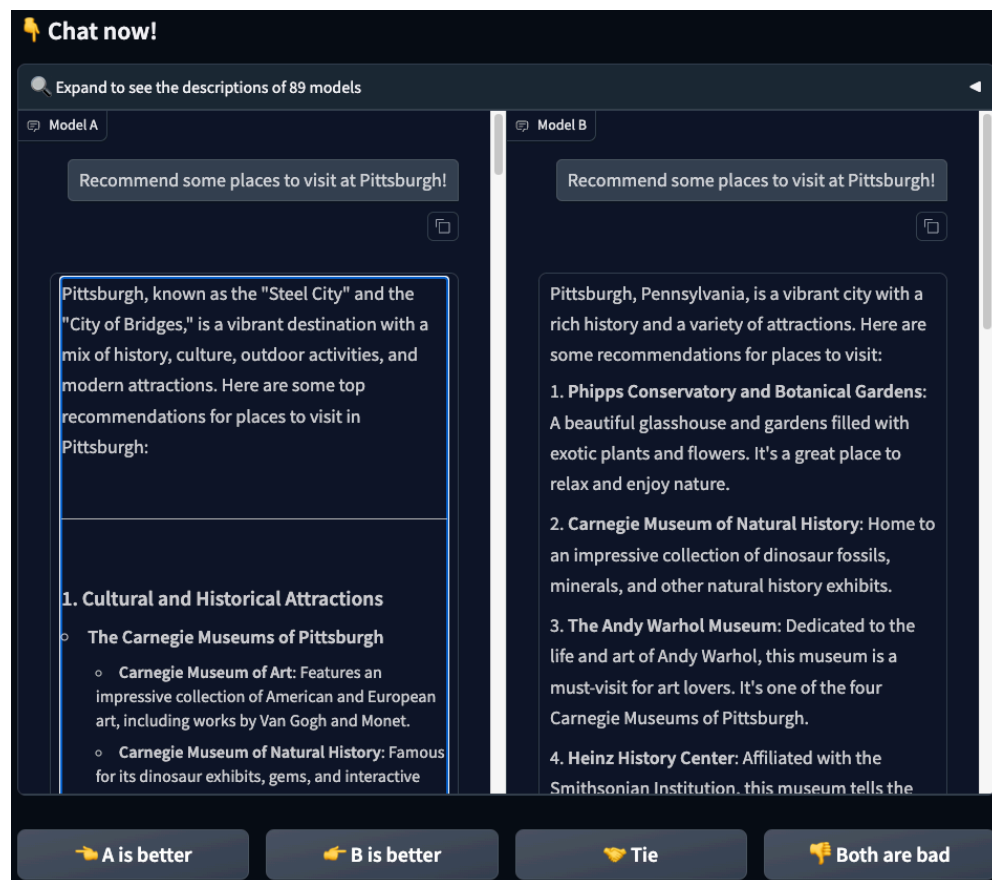
[Lee et al 2025]

- Define criteria $c_1, \dots, c_N$
- Given a response, the LLM judge checks whether each criterion is met



Automatically evaluating complex text

- Option 3: use humans
 - LMSys Chatbot Arena: ask humans to vote which response is better



The screenshot shows the LMSys Chatbot Arena leaderboard. The title is "Leaderboard Overview". Below the title, it says "See how leading models stack up across text, image, vision, and b each Arena, you can explore deeper insights in their dedicated tab". The table shows the top models for the "Text" category, ranked by score. The table has columns for Rank (UB), Model, and Score. The top models are Claude Sonnet 4.5, Gemini 2.5 Pro, Claude Opus 4.1, Claude Sonnet 4.5, and ChatGPT 4o.

Text		
3 days ago		
Rank (UB) ↑	Model ↓	Score ↑↓
1	AI claude-sonnet-4-5-20250929-t...	1453
1	gemini-2.5-pro	1452
1	AI claude-opus-4-1-20250805-thi...	1449
1	AI claude-sonnet-4-5-20250929	1439
2	chatgpt-4o-latest-20250326	1441

Discuss: Pros? Cons?

Discussion: properties of good benchmarks?

- Breadth/diversity
- Depth/difficulty
- Utility
- Robustness
- Data (un-)contamination
- Efficient evaluation (automatic, low cost)

Discussion: properties of good benchmarks?

- Breadth/diversity
- Depth/difficulty
- Utility
- Robustness
- Data (un-)contamination
- Efficient evaluation (automatic, low cost)

Online measurement: product development

- **context**

- industrial settings
- research platforms (e.g., arXiv)

- **evaluators**

- data scientists
- machine learning engineers
- product leadership
- researchers

- **objective**

- corporate objectives
- product objectives
- science

Explicit feedback

what is explicit feedback

View other drafts

Explicit feedback is any feedback that is clearly stated and directly communicated. It is the opposite of implicit feedback, which is inferred from someone's actions or behavior.

Here are some examples of explicit feedback:

- A teacher telling a student that their essay is well-written and provides specific examples of why.
- A customer writing a review of a product, stating what they liked and disliked about it.
- A manager giving an employee feedback on their performance, including both positive and negative feedback.

Explicit feedback can be helpful in a variety of situations, such as:

- **Learning:** Explicit feedback can help people learn and improve their skills. For example, a student who receives feedback on their writing can use that feedback to improve their writing skills in the future.
- **Performance:** Explicit feedback can help people improve their performance at work or school. For example, an employee who receives feedback on their performance can use that feedback to improve their work performance.
- **Relationships:** Explicit feedback can help people improve their relationships with others. For example, a friend who receives feedback from their partner on how they can be more supportive can use that feedback to improve their relationship.

It is important to note that explicit feedback should be given in a constructive way. This means that it should be specific, actionable, and focused on the behavior, not the person. For example, instead of saying "You're a terrible writer," you could say "I think your essay could be improved by adding more specific examples."

I hope this helps!

Check completed • Understand the results

Share, Copy, Print, and other icons

language technology

tecnología del lenguaje

19 / 5,000

Share, Copy, and other icons

Explicit feedback

HA

Please proof read "I am an cute cat. likes milk"

Which response do you prefer?
Your choice will help make ChatGPT better.

 Response 1

Certainly! Here's the corrected version:

"I am a cute cat. I like milk."

I made the following changes:

1. Replaced "an" with "a" to match the noun "cat."
2. Corrected "likes" to "like" for proper verb agreement.

 Response 2

Here is the proofread version of your sentence:

"I am a cute cat. I like milk."

Why implicit feedback

- Explicit feedback disrupts the user experience
 - adds friction to user task completion
 - no clear incentive to provide feedback
- Self-selection bias
 - Users who provide feedback may not be representative of all users
- **Goal:** find behavioral signals correlated with modeling targets that minimally disrupt user experience.
 - historically includes both informal (e.g., clicks as perfect proxies) and formal models (e.g., statistical models)
- Implicit feedback is a model of...
 - ...unobserved properties
 - ...human preferences

Base metrics

- Behavioral metrics are based on logged interaction data
- Select based on relationship to property or delayed reward
 - relationship may require deeper understanding of what metrics mean
- Logging requires system considerations
 - latency
 - responsiveness
 - storage
 - bandwidth
- Consider costs to user
 - bias
 - trust
 - privacy

Clicks

- In web-based technologies, click reflect a user's choice of one item, potentially amongst several competing items
 - most salient in search, recommendation
- natural signal to measure if trying to determine preferences between,
 - competing system decisions (e.g., rankings)
 - competing elements of system decisions (e.g., documents)

Click Logging with a Redirect Server

Cat Food 101: What You Need to Know About Feeding Your Cat

pets.webmd.com/cats/.../cat-food-101-what-you-need-to-know-about...

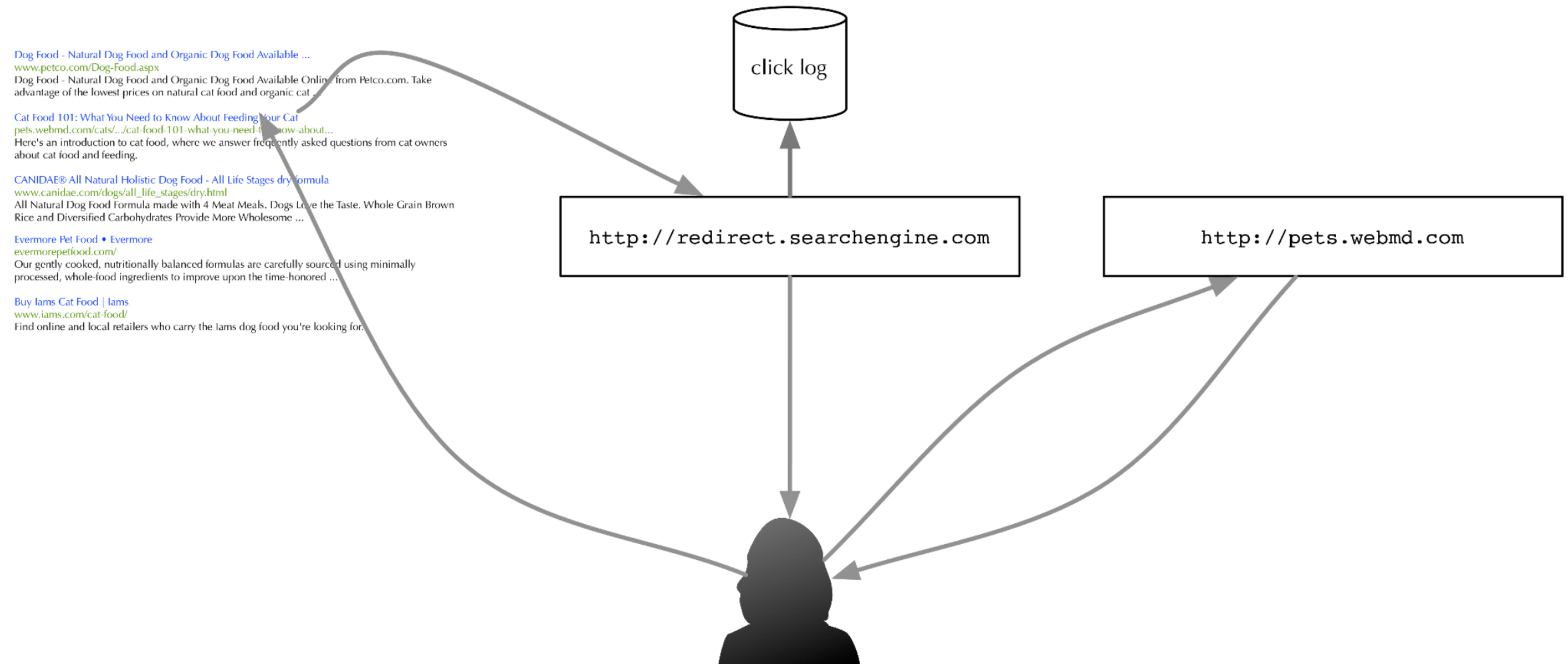
Here's an introduction to cat food, where we answer frequently asked questions from cat owners about cat food and feeding.

```

<div class=res>
  <div class="title"<a href="http://
redirect.searchengine.com/?
sourcedata=A0oG7t6lyBFRZRwAUJvXNyoA&targetdata=http%3a%2f
%2fpets.webmd.com%2fcats%2f...%2fcats-food-101-what-you-need-
to-know-about">Cat Food 101: What You Need to Know About
Feeding Your Cat</a></div>
    <div class="url">pets.webmd.com/
cats/.../cat-food-101-what-you-need-to-know-about...</div>
    <div class="abstract">Here's an
introduction to cat food, where we answer frequently asked
questions from cat owners about cat food and feeding.</div>
  </div>

```

Click Logging with a Redirect Server



Click Logging with a Beacon Server

[Dog Food - Natural Dog Food and Organic Dog Food Available ...](#)

www.petco.com/Dog-Food.aspx

Dog Food - Natural Dog Food and Organic Dog Food Available Online from Petco.com. Take advantage of the lowest prices on natural cat food and organic cat ...

[Cat Food 101: What You Need to Know About Feeding Your Cat](#)

pets.webmd.com/cats/.../cat-food-101-what-you-need-to-know-about...

Here's an introduction to cat food, where we answer frequently asked questions from cat owners about cat food and feeding.

[CANIDAE® All Natural Holistic Dog Food - All Life Stages dry formula](#)

www.canidae.com/dogs/all_life_stages/dry.html

All Natural Dog Food Formula made with 4 Meat Meals. Dogs Love the Taste. Whole Grain Brown Rice and Diversified Carbohydrates Provide More Wholesome ...

[Evermore Pet Food • Evermore](#)

evermorepetfood.com/

Our gently cooked, nutritionally balanced formulas are carefully sourced using minimally processed, whole-food ingredients to improve upon the time-honored ...

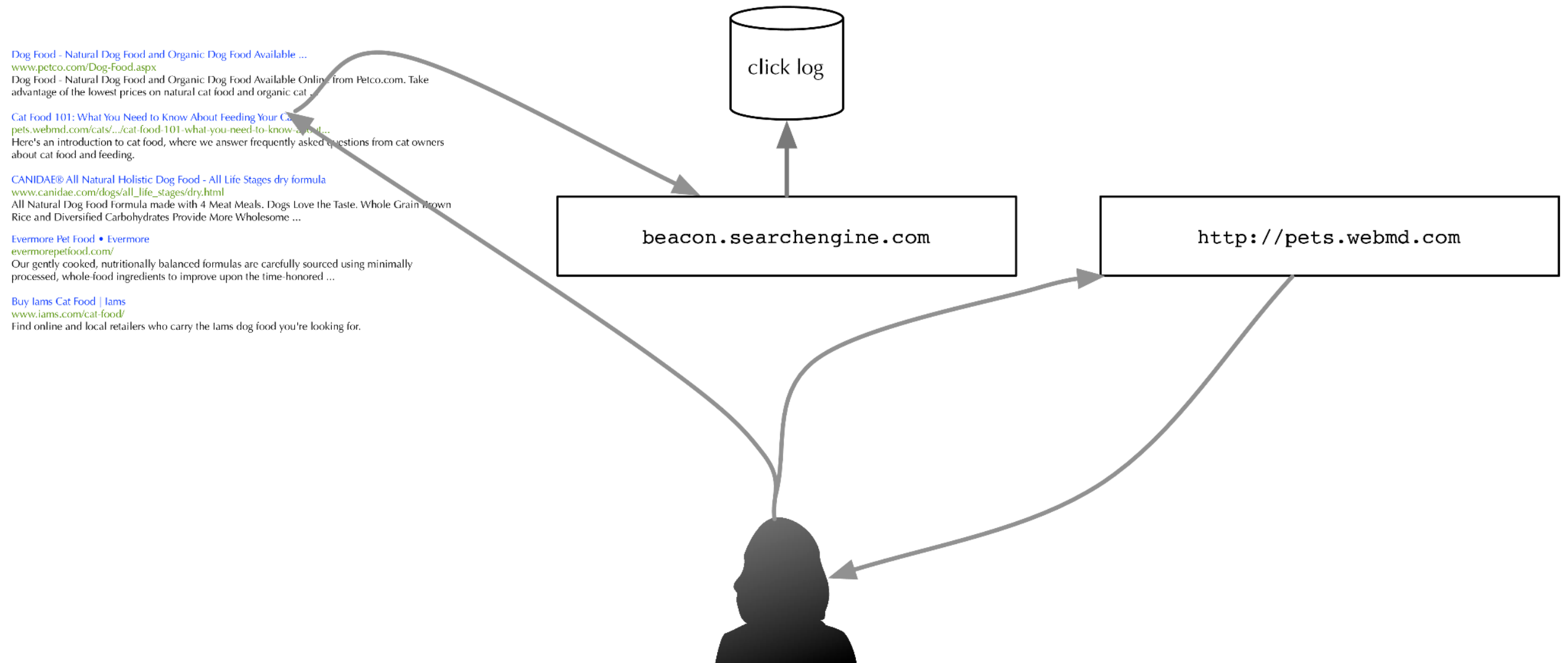
[Buy Iams Cat Food | Iams](#)

www.iams.com/cat-food/

Find online and local retailers who carry the Iams dog food you're looking for.

```
$('#srtitle').delegate('a', 'click', function(event) {  
    var url = this.href;  
    // send source_data and target_data to beacon server  
    window.location = url;  
});
```

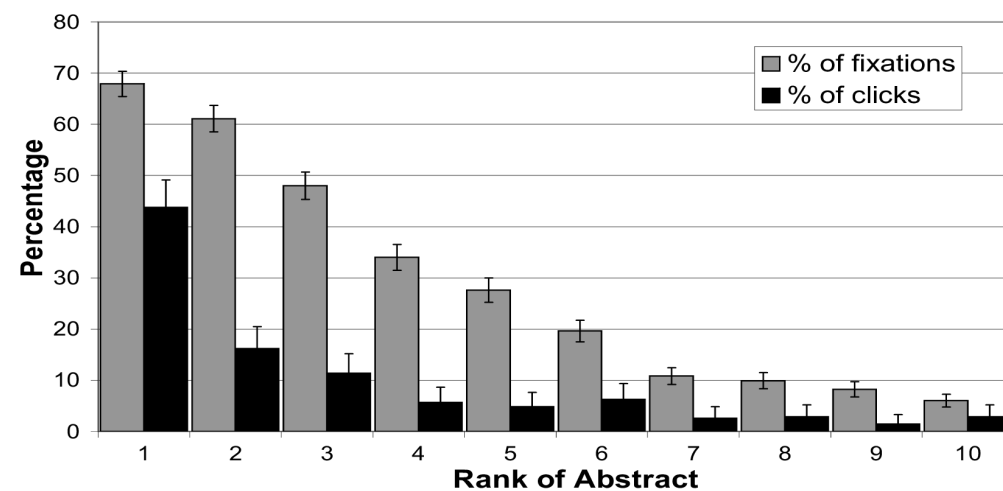
Click Logging with a Beacon Server



Clicks are noisy

- **accidental clicking:** need to detect and remove clicks which immediately return to the search results page.
- **malicious clicking:** need to detect and remove clicks from robots trying to manipulate the system.
- **position bias:** need to account for documents at higher positions being clicked first.
- **presentation bias:** need to account for documents or results that are visually more attractive.
- **ambiguous intent:** need to account for clicks representing different query intents.

Position bias in rankings

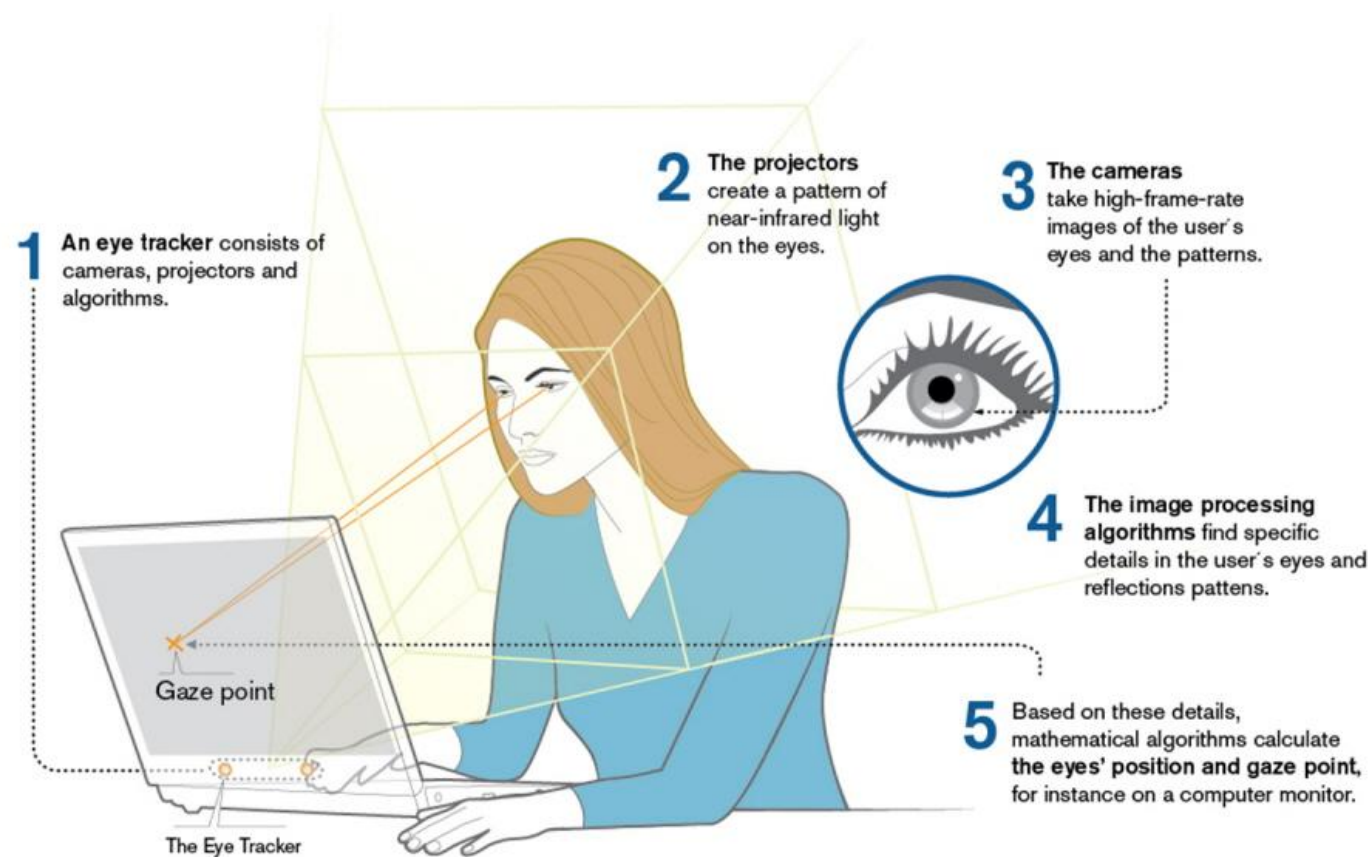


- User scan behavior starts from the top of the ranked list and moves serially downward.
- Documents at the top tend to receive more examinations and clicks.
- Instead of modeling the 'click' or 'no click', we can model clicks conditioned on examination.

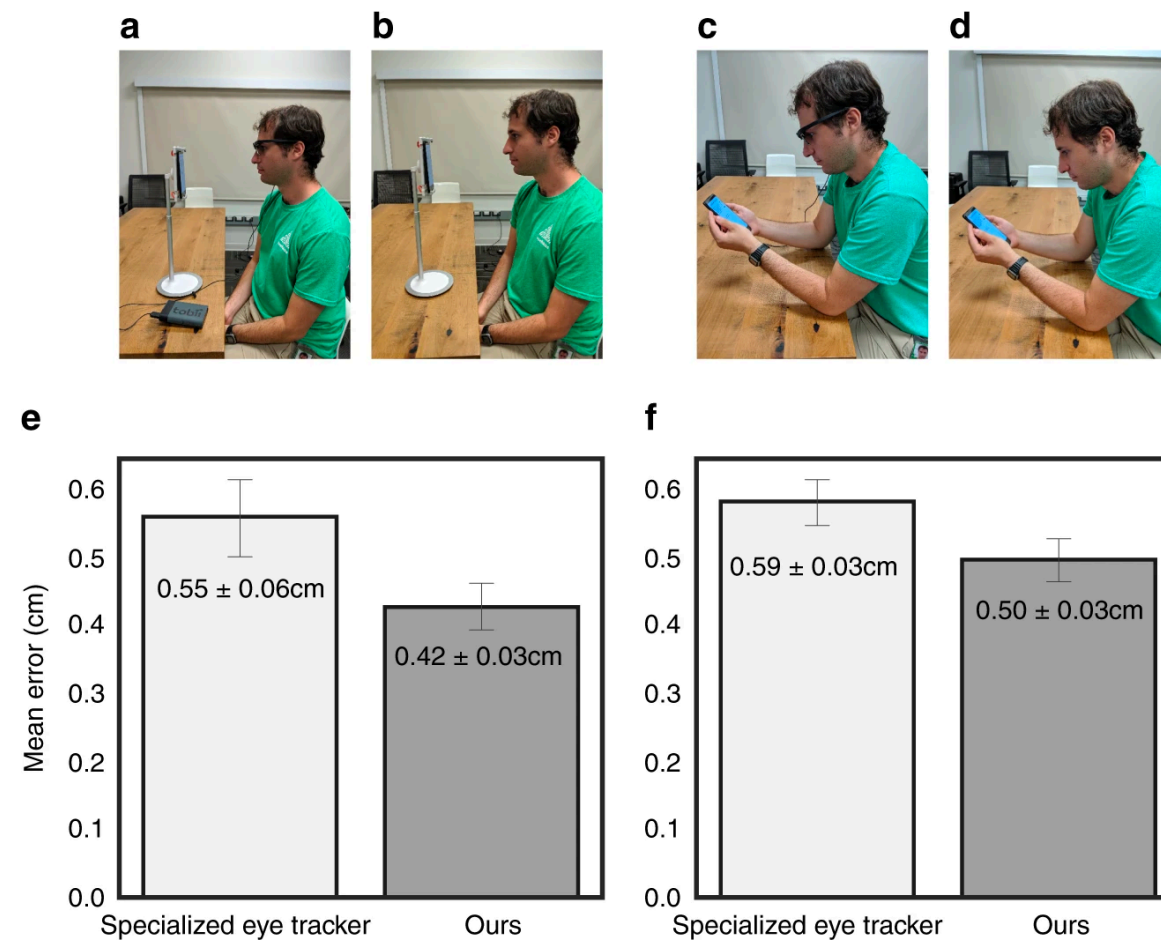
Visual attention

- An item can be *unclicked* for many reasons
 - off-screen
 - on-screen but clicked on something else first
 - not clickable (e.g., word, summary)
- Need a mechanism to log visual attention
 - eye-tracker
 - built-in camera

Visual attention: Eye tracking



Visual attention: Built-in camera



Visual attention: eye fixation correlated with cursor position

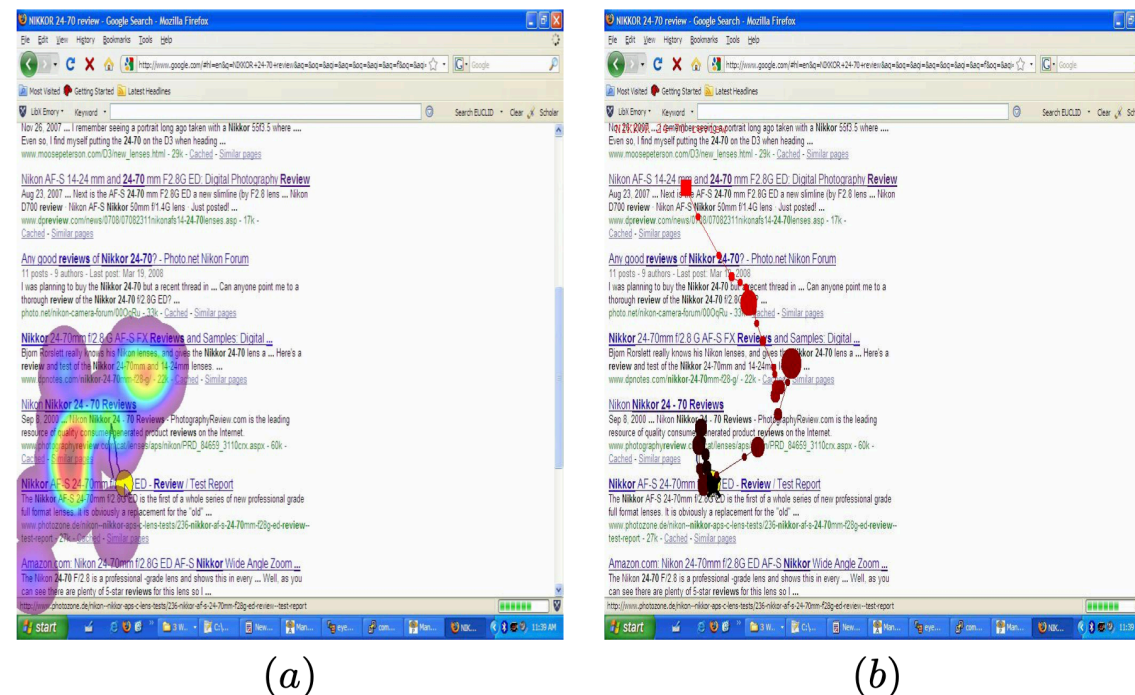


Figure 1: Searcher gaze position and corresponding mouse trajectory (Research intent)

Qi Guo and Eugene Agichtein. Ready to buy or just browsing?: detecting web searcher goals from interaction data. In Proceedings of the 33rd international acm sigir conference on research and development in information retrieval, SIGIR '10, 130--137, New York, NY, USA, 2010. , ACM.

Visual attention: eye fixation correlated with cursor position

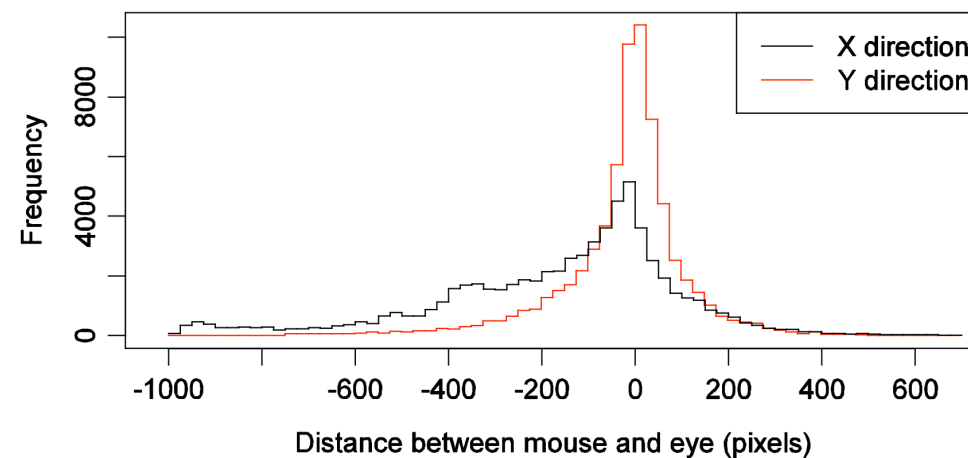


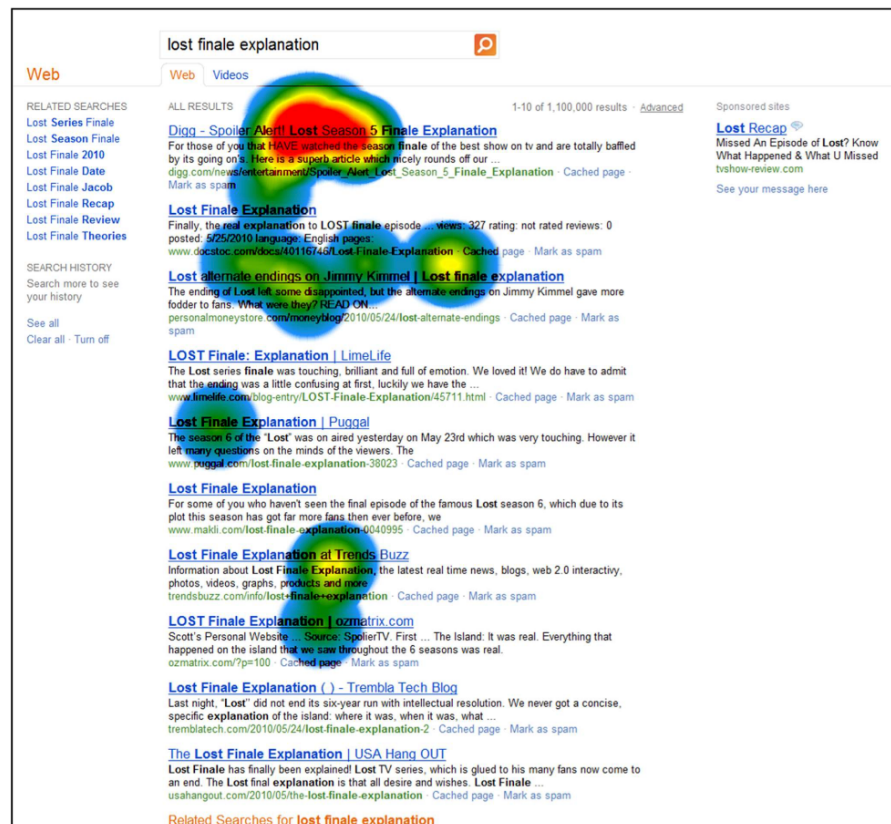
Figure 2: Histogram outlines of distance from mouse to eye, broken out separately for the X and Y directions. Each step in the histogram represents a bucket of 25 pixels. A negative X distance means that the eye was to the left of the mouse, and a negative Y distance means that the eye was below the mouse.

How to Log Cursor Movement

- Instrument the browser.
- Instrument landing pages (e.g. analytics javascript).
- Instrument the results page (e.g. onmousemove event).

Cursor tracking denser than click tracking

Click positions



Cursor movement positions



Figure 2. Heatmaps of all click positions (left) and recorded cursor positions (right) for the query *[lost finale explanation]*. Heavy interaction occurs in red/orange/yellow areas, moderate interaction in green areas, light interaction in blue areas.

Jeff Huang, Ryen W. White, and Susan Dumais. No clicks, no problem: using cursor movements to understand and improve search. In Proceedings of the 2011 annual conference on human factors in computing systems, 1225–1234, 2011.

Viewport tracking



Figure 2. An illustrative example of a heatmap on a Web page generated from mobile device users' viewport data.

Jeff Huang, Ryen W. White, and Susan Dumais. No clicks, no problem: using cursor movements to understand and improve search. In Proceedings of the 2011 annual conference on human factors in computing systems, 1225–1234, 2011.