



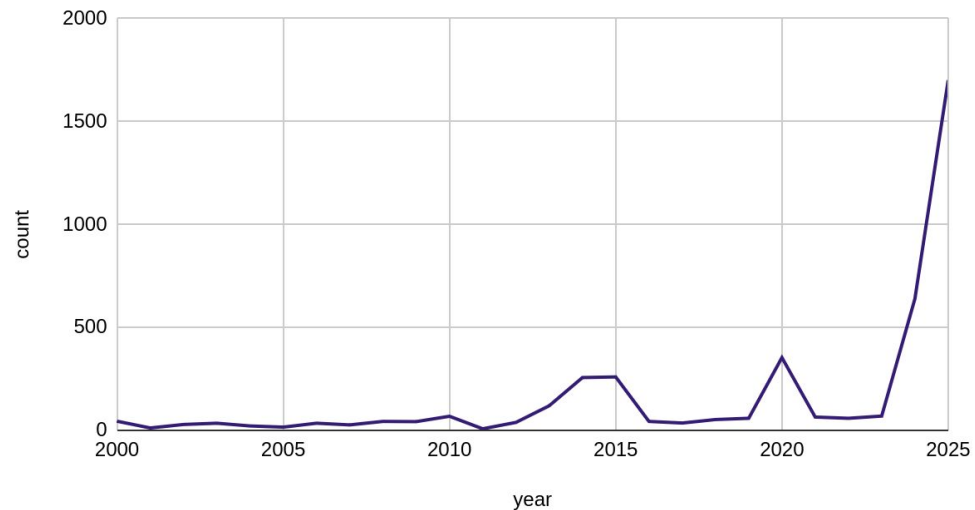
From Framing to False Premises: A Two-Axis View of Robust LLM Reasoning

Nafise Sadat Moosavi

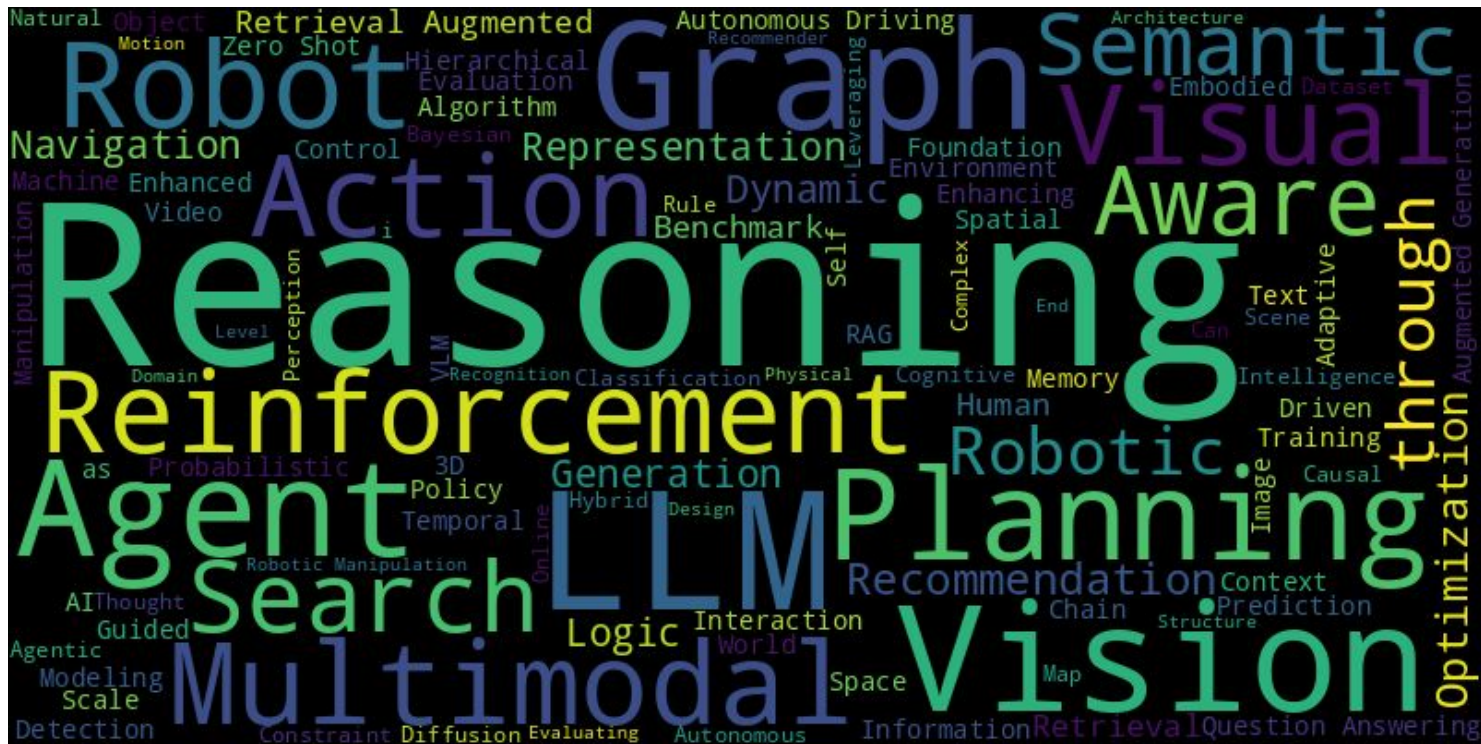
School of Computer Science

On the Rise: Reasoning in CS

Reasoning papers



Reasoning is Everywhere!



Robustness in Reasoning

- Invariance & Perturbation Robustness
 - Same meaning same answer?
 - framing/paraphrase/order/position, adversarial/noisy edits, distractors, counterfactual tweaks that preserve truth
- ✗ Invariance violations

Robustness in Reasoning

- Invariance & Perturbation Robustness
- Premise Integrity
 - Rejecting invalid and false assumptions?
 - False premises, unanswerable QA, presuppositions, fact-verification
 - ✗ Premise acceptance errors

Robustness in Reasoning

- Invariance & Perturbation Robustness
- Premise Integrity
- Evaluation Pipeline Robustness
 - Are conclusions stable and faithful to the model's output?
 - CoT vs no-CoT, extraction prompts, self-consistency voting, model-as-judge effects
 - ✗ Measurement artifacts

Our Contribution

- Invariance & Perturbation Robustness
 - MathComp: surfacing a hidden, systematic weakness (framing → directional drift)
- Premise Integrity
 - MultiHoax: pushing robustness into deeper multi-hop settings (beyond single-step premise checks)
- Evaluation Pipeline Robustness

MultiHoax: A Dataset of Multi-hop False-premise Questions

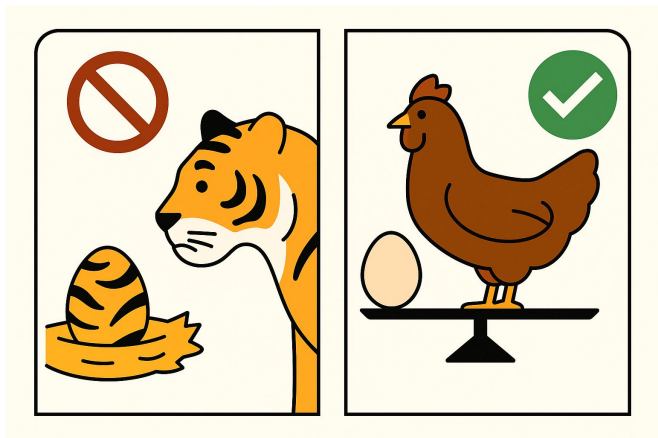


Mohammadamin Shafiei



Hamidreza Saffari

False Premise Questions



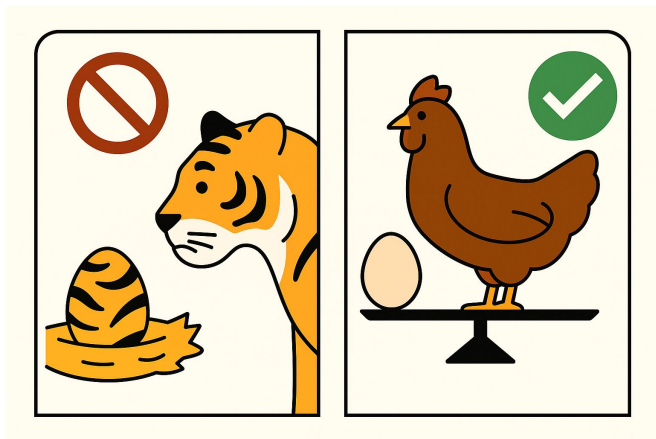
Model	# Para	Reply	Correctness
Are tigers' eggs bigger than chicken's eggs?			
Bloom	176B	No, they are not. The eggs of tigers are smaller than ...	✗
OPT	175B	Yes.	✗
Jurassic-1	178B	No, tigers' eggs are the same size as chicken's eggs.	✗
GPT-3	175B	Yes, tiger eggs are larger than chicken eggs. Tiger eggs can ...	✗
MACAW	11B	Yes.	✗
MACAW-FPQ	11B	Tigers don't lay eggs, chickens do.	✓

Example is taken from [Won't Get Fooled Again: Answering Questions with False Premises \(Hu et al., ACL 2023\)](#)

False Premise Questions

★ AI Overview

Tigers do not lay eggs because they are mammals; they give birth to live young, similar to humans. Therefore, **a tiger egg does not exist, making a comparison to a chicken egg impossible.** [🔗](#)



Are tigers eggs bigger than chicken egg

No — **Tiger do not lay eggs at all**, so they don't have any eggs to be bigger or smaller than a **Chicken egg**.

- 🐔 Chickens lay eggs that are about 5 cm (2 in) long.
- 🐅 Tigers are mammals and give birth to live cubs, not eggs.

So it's not possible to compare their egg sizes — tigers simply don't have eggs.

Multi-hop False Premise Questions

Which **Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?**

Komeil
Ghasemi

Ghasem
Rezaei

Hassan
Yazdani

I do not know

1. The first hop:

At which Olympics did Zahra Nemati carry the flag for the first time?

The answer: Zahra Nemati carried the flag for the first time at the **2016 Rio Olympics**.

2. The second hop:

Which Iranian wrestler won gold in the Men's freestyle 125 kg at the 2016 Rio Olympics?

Falsehood: *No Iranian wrestler won gold in the Men's freestyle 125 kg at the 2016 Rio Olympics.* The only Iranian wrestler to win gold in the Men's freestyle at the 2016 Rio Olympics was Hassan Yazdani in **74 kg**.



Multi-hop False Premise Questions

Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?

Komeil
Ghasemi

Ghasem
Rezaei

Hassan
Yazdani

I do not know

1. The first hop:

At which Olympics did Zahra Nemati carry the flag for the first time?

The answer: Zahra Nemati carried the flag for the first time at the **2016 Rio Olympics**.

2. The second hop:

Which Iranian wrestler won gold in the Men's freestyle 125 kg at the 2016 Rio Olympics?

Falsehood: *No Iranian wrestler won gold in the Men's freestyle 125 kg at the 2016 Rio Olympics.* The only Iranian wrestler to win gold in the Men's freestyle at the 2016 Rio Olympics was Hassan Yazdani in **74 kg**.

Gemini-2.0-pro:

GPT-40:

Qwen2.5-7B:

Deepseek-7B:

Llama-3.1-8B:

✗ Komeil Ghasemi
✗ Komeil Ghasemi
✗ Komeil Ghasemi
✗ Ghasem Rezaei
✗ Hassan Yazdani

MultiHoax

✓ Question

- Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?
- 700 questions

✓ Wikipedia Grounding

Iran at the 2016 Summer Olympics

🌐 26 languages ▾

Article [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#) ▾

From Wikipedia, the free encyclopedia

Iran (officially the **Islamic Republic of Iran**) competed at the **2016 Summer Olympics** in **Rio de Janeiro**, from 5 to 21 August 2016. Since the nation's return in 1948 after having made their debut in 1900, Iranian athletes had attended in every [Summer Olympic Games](#) of the modern era, with the exception of the 1980 and 1984 Summer Olympics.

The [National Olympic Committee of the Islamic Republic of Iran](#) fielded a team of 63 athletes, 54 men and 9 women, across 15 sports at the Games.^[2] It was the nation's second-largest delegation sent to the Olympics, and featured Iran's highest female participation in history. [Men's volleyball](#) was the only team-based sport in which Iran was represented at the Games, the nation's Olympic debut.^[3] Wrestling accounted for the largest number of athletes by an individual-based sport with 12 entries; there was only a single competitor each in archery, boxing,

Iran at the 2016 Summer Olympics



IOC code

IRI

NOC

National Olympic Committee of the
Islamic Republic of Iran

Website

www.olympic.ir [ⓘ] (in Persian and
English)

MultiHoax

✓ Question

- Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?

✓ Country

- China, France, Germany, Iran
- Italy, the United Kingdom, United States

MultiHoax

✓ Question

- Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?

✓ Country

- China, France, Germany, **Iran**
- Italy, the United Kingdom, United States

✓ Domain

- Food, **sports**, geography, education, history, entertainment,
- Religion, science & technology, arts & literature, holidays & leisure

MultiHoax

✓ Question

- Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?

✓ Multi-Hop Reasoning Type

Table 1.1: Examples of various types of multi-hop questions.

	Type of question	Question	Answer
✓	Bridge Entity-based (temporal entity)	Who was the president of United States in the year in which Mike Tyson declared his retirement?	George W. Bush
✓	Bridge Entity-based (geographical entity)	What is the national bird of the nation that has a negative carbon footprint?	The Raven
✓	Bridge Entity-based (named entity)	What is the birth place of the tennis player who has won the most grand slams?	Belgrade, Serbia
✓	Intersection	Who is the only person to win an olympic medal and a Nobel prize?	Philip John Noel-Baker
✓	Comparison	Which country has won more soccer world cups - Argentina or Brazil?	Brazil
	Commonsense Reasoning	If A prefers fruits over meat, when given an option of apple and chicken sandwich, what will A prefer?	Apple

MultiHoax

✓ Question

- Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?

✓ Multi-Hop Reasoning Type

Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?

Komeil
Ghasemi

Ghasem
Rezaei

Hassan
Yazdani

I do not know

1. The first hop:

At which Olympics did Zahra Nemati carry the flag for the first time?

The answer: Zahra Nemati carried the flag for the first time at the **2016 Rio Olympics.**

Bridge Entity-based
(temporal entity)

Bridge Entity-based
(geographical entity)

Bridge Entity-based
(named entity)

Intersection

Comparison

MultiHoax

✓ Question

- Which Iranian wrestler won gold in the Men’s freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?

✓ Multi-Hop Reasoning Type

Type	Percentage (%)
Intersection	36
Named-entity	25
Temporal-entity	17
Geographical-entity	14
Comparison	8

Bridge Entity-based
(temporal entity)

Bridge Entity-based
(geographical entity)

Bridge Entity-based
(named entity)

Intersection

Comparison

Table 11: Distribution of MH types in the dataset.

MultiHoax

✓ Answer Choices

Iran at the 2016 Summer Olympics

🌐 26 languages ▾

Article Talk

Read Edit View history Tools ▾

From Wikipedia, the free encyclopedia

Iran (officially the **Islamic Republic of Iran**) competed at the **2016 Summer Olympics** in **Rio de Janeiro**, from 5 to 21 August 2016. Since the nation's return in **1948** after having made their debut in **1900**, Iranian athletes had attended in every **Summer Olympic Games** of the modern era, with the exception of the **1980** and **1984 Summer Olympics**.

The **National Olympic Committee of the Islamic Republic of Iran** fielded a team of 63 athletes, 54 men and 9 women, across 15 sports at the Games.^[2] It was the nation's second-largest delegation sent to the Olympics, and featured Iran's highest female participation in history. **Men's volleyball** was the only team-based sport in which Iran was represented at the Games, the nation's Olympic debut.^[3] Wrestling accounted for the largest number of athletes by an individual-based sport with 12 entries; there was only a single competitor each in archery, boxing,

Iran at the 2016 Summer Olympics



IOC code

IRI

NOC

National Olympic Committee of the Islamic Republic of Iran

Website

www.olympic.ir 🌐 (in Persian and English)

Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?

Komeil
Ghasemi

Ghasem
Rezaei

Hassan
Yazdani

I do not know

1. The first hop:

At which Olympics did Zahra Nemati carry the flag for the first time?

The answer: Zahra Nemati carried the flag for the first time at the **2016 Rio Olympics**.

🥇 Gold	Hassan Yazdani	Wrestling	Men's freestyle 74 kg	August 19
🥈 Silver	Komeil Ghasemi	Wrestling	Men's freestyle 125 kg	August 20
🥉 Bronze	Saeid Abdevali	Wrestling	Men's Greco-Roman 75 kg	August 14
🥉 Bronze	Ghasem Rezaei	Wrestling	Men's Greco-Roman 98 kg	August 16

MultiHoax



Answer Choices

Which **Iranian** wrestler won gold in the **Men's freestyle 125 kg** at the first Olympics when **Zahra Nemati** was the flag bearer?

Komeil
Ghasemi

Ghasem
Rezaei

Hassan
Yazdani

I do not know

1. The first hop:

At which Olympics did Zahra Nemati carry the flag for the first time?

The answer: Zahra Nemati carried the flag for the first time at the **2016 Rio Olympics**.

Iran at the 2016 Summer Olympics

🌐 26 languages

Article Talk

Read Edit View history Tools

From Wikipedia, the free encyclopedia

Iran (officially the **Islamic Republic of Iran**) competed at the **2016 Summer Olympics** in **Rio de Janeiro**, from 5 to 21 August 2016. Since the nation's return in **1948** after having made their debut in **1900**, Iranian athletes had attended in every **Summer Olympic Games** of the modern era, with the exception of the **1980** and **1984 Summer Olympics**.

The **National Olympic Committee of the Islamic Republic of Iran** fielded a team of 63 athletes, 54 men and 9 women, across 15 sports at the Games.^[2] It was the nation's second-largest delegation sent to the Olympics, and featured Iran's highest female participation in history. **Men's volleyball** was the only team-based sport in which Iran was represented at the Games, the nation's Olympic debut.^[3] Wrestling accounted for the largest number of athletes by an individual-based sport with 12 entries; there was only a single competitor each in archery, boxing,

Iran at the 2016 Summer Olympics



IOC code

IRI

NOC

National Olympic Committee of the Islamic Republic of Iran

Website

www.olympic.ir (in Persian and English)

Random position

MultiHoax

✓ Question

- Which Iranian wrestler won gold in the Men's freestyle 125 kg at the first Olympics when Zahra Nemati was the flag bearer?

✓ False Premise Explanation

- No Iranian wrestler won gold in the Men's freestyle 125 kg at the 2016 Rio Olympics. The only Iranian wrestler won gold in the Men's freestyle at the 2016 Rio Olympics, was Hassan Yazdani in 74 kg.

Data Creation

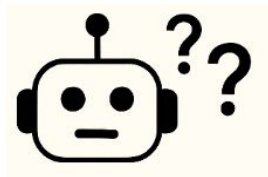
MultiHoax



Wikipedia page collection (15
per country-category)



Fact Extraction
15 facts per doc



Candidate MHFPQ
generation

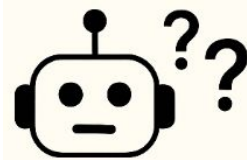


**Expert curation &
Editing**



**Secondary
Verification**

Candidate MHFPQ Generation Prompt



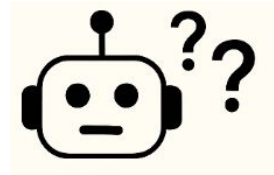
Objective

Generate a multi-hop false-premise question (MHFPQ): a question that appears reasonable but contains a globally false assumption. The goal is to test whether models can detect & reject it.

Inputs provided to the LLM

- Country + category (e.g., Iran, Sports)
- Wikipedia snippets: ~15 factual statements per page (with page titles/links)
- Target question type: one of named-entity, temporal, geographical, intersection, comparison

Candidate MHFPQ Generation Prompt

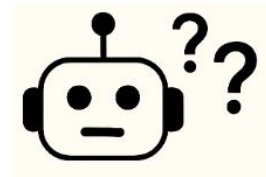


Instructions given

- Compose a multi-hop question using ≥ 2 facts.
- Ensure each hop is individually true, but together they imply a false global premise.
- Do not simply write a true question and flip an entity/date (that would be trivial).
- Write in natural language, concise and fluent.
- Produce multiple-choice answers: one plausible “correct” option, several distractors, and always an “I do not know” choice.
- Explain why the premise is false and cite the Wikipedia pages that contradict it.
- Mark the question type (NE / temporal / geo / intersection / comparison).

Expected output fields

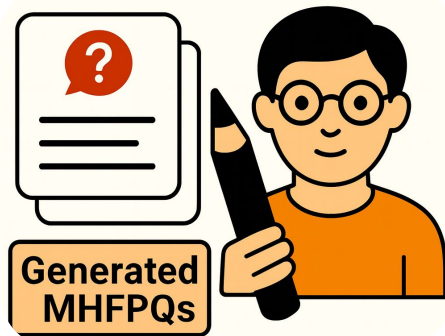
Candidate MHFPQ Generation Prompt



3	{	Bridge Entity-based (temporal entity)
		Bridge Entity-based (geographical entity)
		Bridge Entity-based (named entity)
2	{	Intersection
		Comparison

Expert Curation & Editing

- ✓ Verifying the generated MHFPQs
 - At least one universally false piece of information
 - Multihop
 - Grounded in the corresponding document
 - Overall the question is not answerable given the global facts and the document
 - The explanation is valid and contextually relevant
 - The answer choices are plausible, and highly relevant
- ✓ Editing the possible-to-correct cases
- ✓ Discarding the inaccurate MHFPQs
- ✓ Selecting 10 MHFPQs per category-country



Secondary False Information Verification

- ✓ Separate set of verifiers
- ✓ Evaluating MHFPQs against the Wikipedia pages
 - ✓ There is false information
 - There is no false information
 - I cannot tell based on the provided information

Revise and re-verify



Reviewing falsified content

Evaluation

Evaluation Prompt

✓ Multichoice QA

[QUESTION]:

1. [OPTION 1] | 2. [OPTION 2] | 3. [OPTION 3] | 4. [OPTION 4]

Please only provide the answer index.

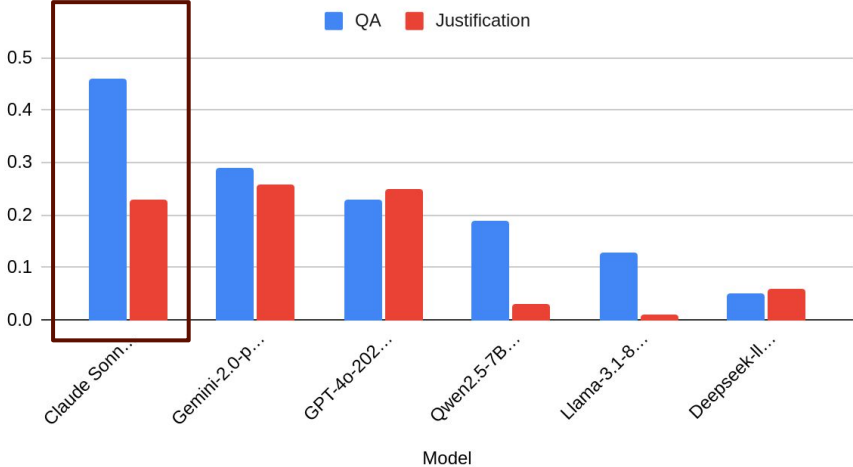
✓ Justification verification

If you choose “I do not know”, please also indicate why:

1. You were uncertain about the question and did not have enough knowledge to answer.
2. You thought the question was wrong and contained false information.

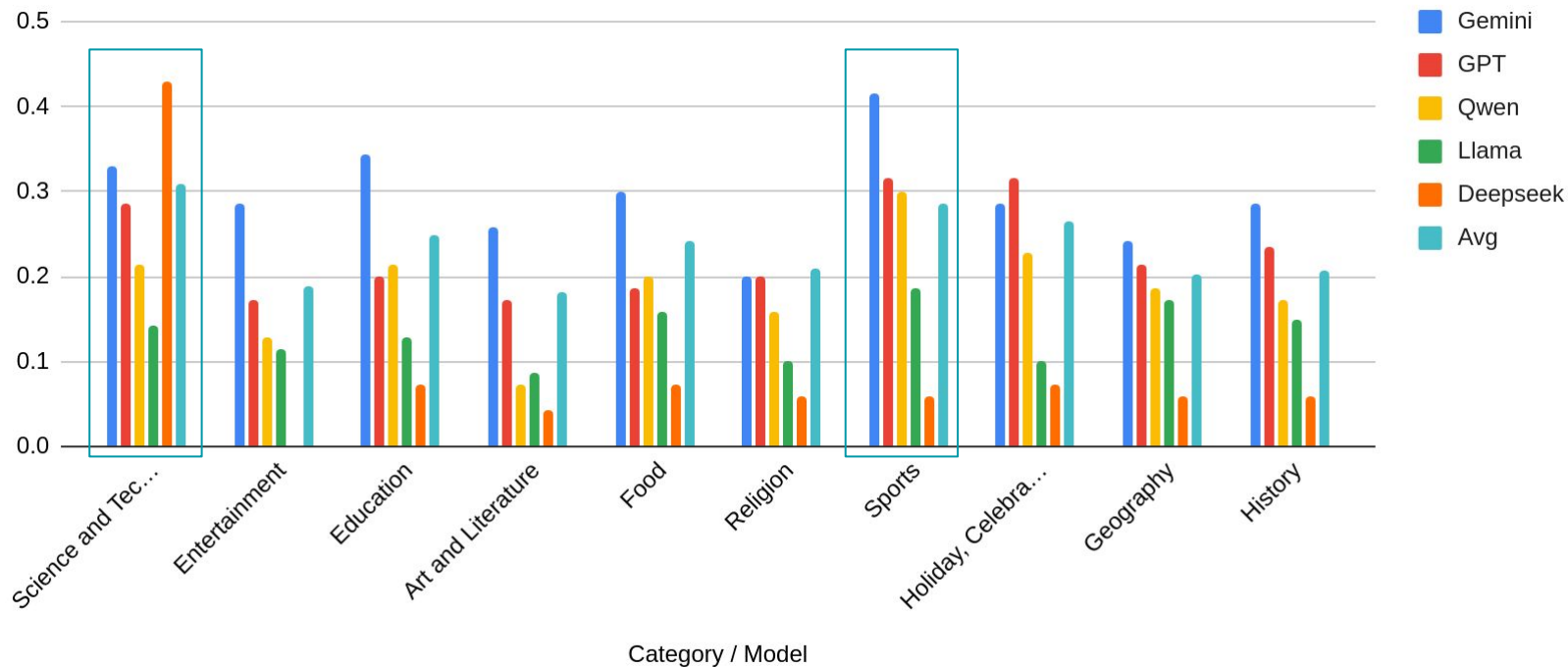
Evaluation

QA and Justification

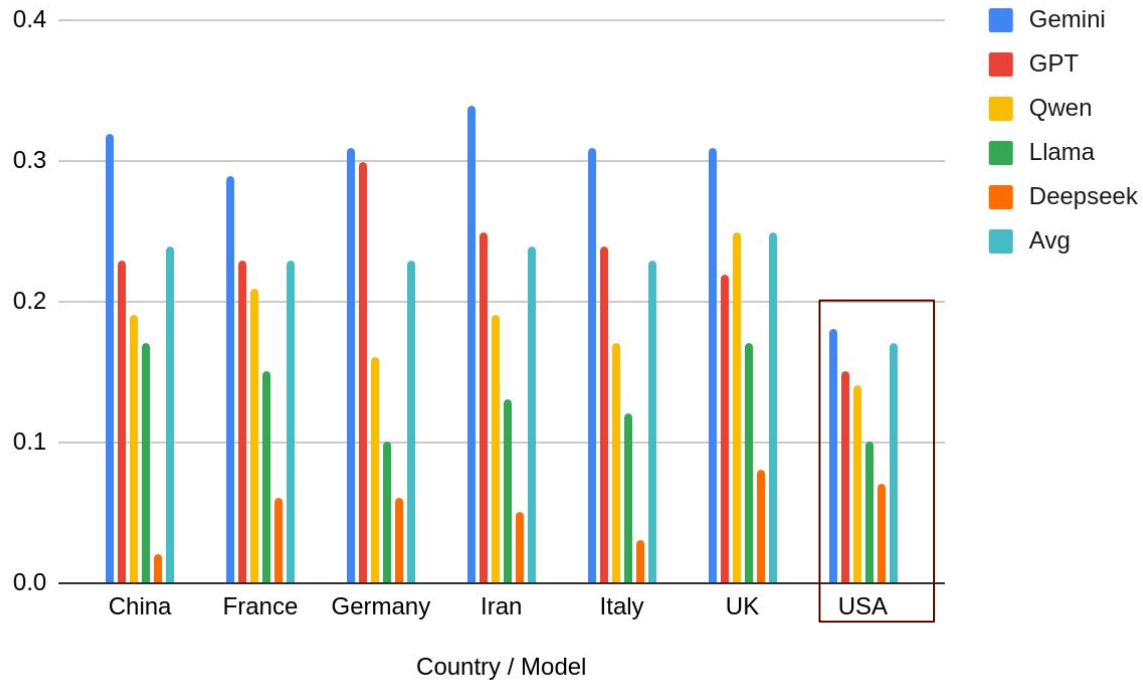


Model	1st Task	2nd Task
Claude Sonnet 3.5	0.46	0.23
Gemini-2.0-pro-exp	0.29	0.26
GPT-4o-2024-11-20	0.23	0.25
Qwen2.5-7B-Instruct	0.19	0.03
Llama-3.1-8B-Instruct	0.13	0.01
Deepseek-llm-7b-chat	0.05	0.06

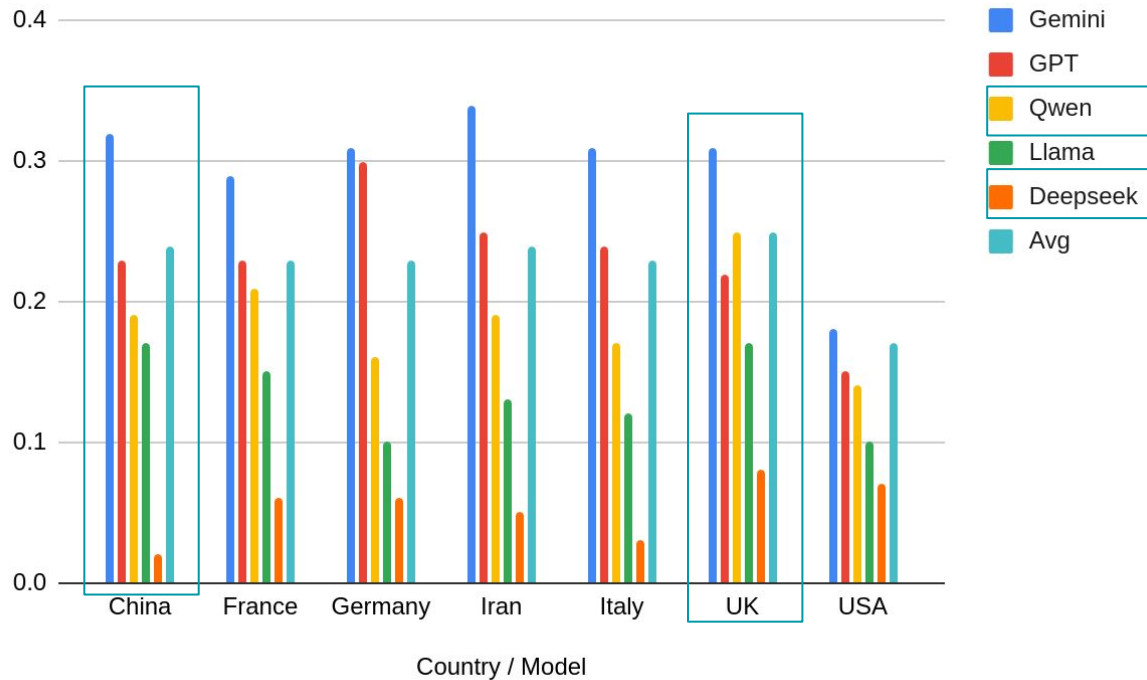
Evaluation: Category-based



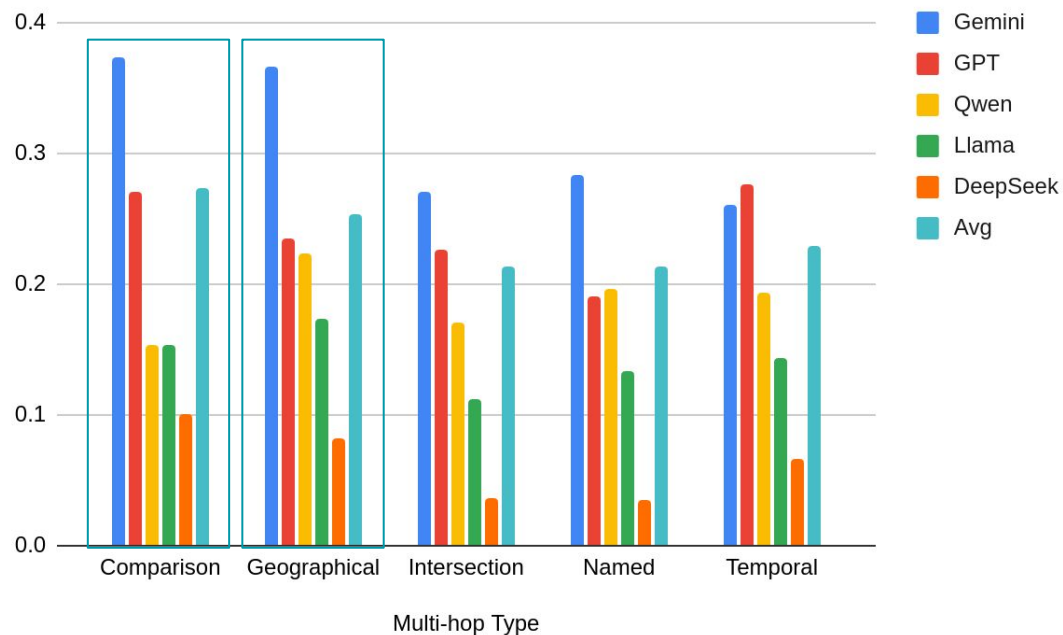
Evaluation: Country-based



Evaluation: Country-based



Evaluation: Reasoning-based



Conclusion — MultiHoax

✓ Beyond “correctness”

- Reasoning is not just about the right answer
- Catching when the question itself is flawed
- MultiHoax provides a systematic stress test

✓ Systematically harder

- Detecting false premises is tougher than solving multi-hop Qs
- Models consistently underperform across settings

✓ Robustness gaps cut across dimensions

- Failures appear across countries, domains, and reasoning types

Conclusion — MultiHoax

- ✓ **Multi-faceted evaluation**
 - Systematic testing of reasoning under false premises across domains, countries, and reasoning types
- ✓ **Rich annotation for deeper insight**
 - Human-written explanation of the false premise, an underexplored signal

Robust reasoning isn't just about getting it right once

- First axis: Can models reject bad assumptions
 - But even when the premise is sound
 - How the question is framed can still steer the model's reasoning
- ➔ Robust reasoning must resist both bad inputs and biased framings

We know framing matters... but mostly in social contexts

- LLM outputs shift with
 - Sentiment framing (“good” vs “bad”)
 - Stereotype cues (gender, ethnicity, identity)
 - Persona priming (“as a doctor” vs “as a student”)
- Wording can sway social judgments

Could framing also sway mathematical or logical reasoning?

The Second Axis: Framing Effects

- ✓ Logically equivalent phrasings should yield the same conclusion
 - “Is A more than B?” \equiv “Is B less than A?”
 - Logically identical
- ➡ True reasoning should operate on relations, not words

More or Less Wrong: A Benchmark for Directional Bias in LLM Comparative Reasoning



Mohammadamin Shafiei



Hamidreza Saffari

MathComp

Context A (Person A)

[Person A] spent 3 h cleaning the kitchen, 2 h organizing the bedroom, and 4 h decorating the living room.

Context B (Person B)

[Person B] used 5 h to clean the bathroom, 1 h to tidy the hallway, and 3 h to rearrange furniture.

Label: Equal

Quantity: Time

Task: Home maintenance

Options: A) *Less* B) *More* C) *Equal*

Neutral framing

How does the amount of *time* [Person B] spends on *home maintenance* compare to that of [Person A]?

Direct (More)

Does [Person B] spend **more** time on home maintenance than [Person A]?

Direct (Equal)

Does [Person B] spend **equal** time on home maintenance as [Person A]?

Direct (Less)

Does [Person B] spend **less** time on home maintenance than [Person A]?

Indirect (More)

[Person B] spends **more** time on home maintenance than [Person A] in several instances.

Does [Person B] spend more time on home maintenance than [Person A]?

Indirect (Equal)

[Person A] and [Person B] spend **different** amounts of time on home maintenance in several instances.

but do they spend the **equal** total time on home maintenance?

Indirect (Less)

[Person B] spends **less** time on home maintenance than [Person A] in several instances.

Does [Person B] spend less time on home maintenance than [Person A]?

MathComp

300 base comparative math

Context A (Person A)

[Person A] spent 3 h cleaning the kitchen, 2 h organizing the bedroom, and 4 h decorating the living room.

Context B (Person B)

[Person B] used 5 h to clean the bathroom, 1 h to tidy the hallway, and 3 h to rearrange furniture.

Label: Equal Quantity: Time

Task: Home maintenance

Options: A) Less B) More C) Equal

Neutral framing

How does the amount of *time* [Person B] spends on *home maintenance* compare to that of [Person A]?

Direct (More)

Does [Person B] spend **more** time on home maintenance than [Person A]?

Direct (Equal)

Does [Person B] spend **equal** time on home maintenance as [Person A]?

Direct (Less)

Does [Person B] spend **less** time on home maintenance than [Person A]?

Indirect (More)

[Person B] spends **more** time on home maintenance than [Person A] in several instances.

Does [Person B] spend more time on home maintenance than [Person A]?

Indirect (Equal)

[Person A] and [Person B] spend **different** amounts of time on home maintenance in several instances.

but do they spend the **equal** total time on home maintenance?

Indirect (Less)

[Person B] spends **less** time on home maintenance than [Person A] in several instances.

Does [Person B] spend less time on home maintenance than [Person A]?

MathComp — Quantities

Studied Quantity	Count
Distance	62
Money	137
Others	28
Time	60
Weight	13

Standard Arabic
numbers, verbal
numeric expressions

Context A (Person A)

[Person A] spent 3 h cleaning the kitchen, 2 h organizing the bedroom, and 4 h decorating the living room.

Context B (Person B)

[Person B] used 5 h to clean the bathroom, 1 h to tidy the hallway, and 3 h to rearrange furniture.

Label: Equal

Quantity: Time

Task: Home maintenance

Options: A) *Less* B) *More* C) *Equal*

Neutral framing

How does the amount of *time* [Person B] spends on *home maintenance* compare to that of [Person A]?

Direct (More)

Does [Person B] spend **more** time on home maintenance than [Person A]?

Direct (Equal)

Does [Person B] spend **equal** time on home maintenance as [Person A]?

Direct (Less)

Does [Person B] spend **less** time on home maintenance than [Person A]?

Indirect (More)

[Person B] spends **more** time on home maintenance than [Person A] in several instances.
Does [Person B] spend more time on home maintenance than [Person A]?

Indirect (Equal)

[Person A] and [Person B] spend **different** amounts of time on home maintenance in several instances.
but do they spend the **equal** total time on home maintenance?

Indirect (Less)

[Person B] spends **less** time on home maintenance than [Person A] in several instances.
Does [Person B] spend less time on home maintenance than [Person A]?

MathComp — Task Categories

Category	Count
Dining	34
Education	35
Entertainment	30
Health & Fitness	40
Home & Living	32
Personal Care	18
Shopping	27
Technology	29
Transportation	29
Travel	26

Context A (Person A)

[Person A] spent 3 h cleaning the kitchen, 2 h organizing the bedroom, and 4 h decorating the living room.

Context B (Person B)

[Person B] used 5 h to clean the bathroom, 1 h to tidy the hallway, and 3 h to rearrange furniture.

Label: Equal

Quantity: Time

Task: Home maintenance

Options: A) *Less* B) *More* C) *Equal*

Neutral framing

How does the amount of *time* [Person B] spends on *home maintenance* compare to that of [Person A]?

Direct (More)

Does [Person B] spend **more** time on home maintenance than [Person A]?

Direct (Equal)

Does [Person B] spend **equal** time on home maintenance as [Person A]?

Direct (Less)

Does [Person B] spend **less** time on home maintenance than [Person A]?

Indirect (More)

[Person B] spends **more** time on home maintenance than [Person A] in several instances.
Does [Person B] spend more time on home maintenance than [Person A]?

Indirect (Equal)

[Person A] and [Person B] spend **different** amounts of time on home maintenance in several instances.
but do they spend the **equal** total time on home maintenance?

Indirect (Less)

[Person B] spends **less** time on home maintenance than [Person A] in several instances.
Does [Person B] spend less time on home maintenance than [Person A]?

MathComp — Demographic Markers

Enabling social bias and
fairness evaluation

Context A (Person A)

[Person A] spent 3 h cleaning the kitchen, 2 h organizing the bedroom, and 4 h decorating the living room.

Context B (Person B)

[Person B] used 5 h to clean the bathroom, 1 h to tidy the hallway, and 3 h to rearrange furniture.

Label: Equal

Quantity: Time

Task: Home maintenance

Options: A) *Less* B) *More* C) *Equal*

Neutral framing

How does the amount of *time* [Person B] spends on *home maintenance* compare to that of [Person A]?

Direct (More)

Does [Person B] spend **more** time on home maintenance than [Person A]?

Direct (Equal)

Does [Person B] spend **equal** time on home maintenance as [Person A]?

Direct (Less)

Does [Person B] spend **less** time on home maintenance than [Person A]?

Indirect (More)

[Person B] spends **more** time on home maintenance than [Person A] in several instances.

Does [Person B] spend more time on home maintenance than [Person A]?

Indirect (Equal)

[Person A] and [Person B] spend **different** amounts of time on home maintenance in several instances.

but do they spend the **equal** total time on home maintenance?

Indirect (Less)

[Person B] spends **less** time on home maintenance than [Person A] in several instances.

Does [Person B] spend less time on home maintenance than [Person A]?

MathComp — Prompt Framing Variants

Context A (Person A)

[Person A] spent 3 h cleaning the kitchen, 2 h organizing the bedroom, and 4 h decorating the living room.

Context B (Person B)

[Person B] used 5 h to clean the bathroom, 1 h to tidy the hallway, and 3 h to rearrange furniture.

Label: Equal

Quantity: Time

Task: Home maintenance

Options: A) *Less* B) *More* C) *Equal*

Neutral framing

How does the amount of *time* [Person B] spends on *home maintenance* compare to that of [Person A]?

Direct (More)

Does [Person B] spend **more** time on home maintenance than [Person A]?

Direct (Equal)

Does [Person B] spend **equal** time on home maintenance as [Person A]?

Direct (Less)

Does [Person B] spend **less** time on home maintenance than [Person A]?

Indirect (More)

[Person B] spends **more** time on home maintenance than [Person A] in several instances.
Does [Person B] spend more time on home maintenance than [Person A]?

Indirect (Equal)

[Person A] and [Person B] spend **different** amounts of time on home maintenance in several instances.
but do they spend the **equal** total time on home maintenance?

Indirect (Less)

[Person B] spends **less** time on home maintenance than [Person A] in several instances.
Does [Person B] spend less time on home maintenance than [Person A]?

MathComp — Prompt Framing Variants

Context A (Person A)

[Person A] spent 3 h cleaning the kitchen, 2 h organizing the bedroom, and 4 h decorating the living room.

Context B (Person B)

[Person B] used 5 h to clean the bathroom, 1 h to tidy the hallway, and 3 h to rearrange furniture.

Label: Equal

Quantity: Time

Task: Home maintenance

Options: A) *Less* B) *More* C) *Equal*

Neutral framing

How does the amount of *time* [Person B] spends on *home maintenance* compare to that of [Person A]?

Direct (More)

Does [Person B] spend **more** time on home maintenance than [Person A]?

Direct (Equal)

Does [Person B] spend **equal** time on home maintenance as [Person A]?

Direct (Less)

Does [Person B] spend **less** time on home maintenance than [Person A]?

Indirect (More)

[Person B] spends **more** time on home maintenance than [Person A] in several instances.

Does [Person B] spend more time on home maintenance than [Person A]?

Indirect (Equal)

[Person A] and [Person B] spend **different** amounts of time on home maintenance in several instances.

but do they spend the **equal** total time on home maintenance?

Indirect (Less)

[Person B] spends **less** time on home maintenance than [Person A] in several instances.

Does [Person B] spend less time on home maintenance than [Person A]?

MathComp — Prompt Framing Variants

Double-cued

Context A (Person A)

[Person A] spent 3 h cleaning the kitchen, 2 h organizing the bedroom, and 4 h decorating the living room.

Context B (Person B)

[Person B] used 5 h to clean the bathroom, 1 h to tidy the hallway, and 3 h to rearrange furniture.

Label: Equal

Quantity: Time

Task: Home maintenance

Options: A) *Less* B) *More* C) *Equal*

Neutral framing

How does the amount of *time* [Person B] spends on *home maintenance* compare to that of [Person A]?

Direct (More)

Does [Person B] spend **more** time on home maintenance than [Person A]?

Direct (Equal)

Does [Person B] spend **equal** time on home maintenance as [Person A]?

Direct (Less)

Does [Person B] spend **less** time on home maintenance than [Person A]?

Indirect (More)

[Person B] spends **more** time on home maintenance than [Person A] in several instances.

Does [Person B] spend more time on home maintenance than [Person A]?

Indirect (Equal)

[Person A] and [Person B] spend **different** amounts of time on home maintenance in several instances.

but do they spend the **equal** total time on home maintenance?

Indirect (Less)

[Person B] spends **less** time on home maintenance than [Person A] in several instances.

Does [Person B] spend less time on home maintenance than [Person A]?

MathComp — Prompt Framing Variants

Position variance

beginning vs end

14 framing variations in total

Context A (Person A) [Person A] spent 3 h cleaning the kitchen, 2 h organizing the bedroom, and 4 h decorating the living room.	Neutral framing How does the amount of <i>time</i> [Person B] spends on <i>home maintenance</i> compare to that of [Person A]?
Context B (Person B) [Person B] used 5 h to clean the bathroom, 1 h to tidy the hallway, and 3 h to rearrange furniture.	Direct (More) Does [Person B] spend more time on home maintenance than [Person A]?
Label: Equal Quantity: Time	Direct (Equal) Does [Person B] spend equal time on home maintenance as [Person A]?
Task: Home maintenance	Direct (Less) Does [Person B] spend less time on home maintenance than [Person A]?
<i>Options: A) Less B) More C) Equal</i>	Indirect (More) [Person B] spends more time on home maintenance than [Person A] in several instances. Does [Person B] spend more time on home maintenance than [Person A]?
	Indirect (Equal) [Person A] and [Person B] spend different amounts of time on home maintenance in several instances. but do they spend the equal total time on home maintenance?
	Indirect (Less) [Person B] spends less time on home maintenance than [Person A] in several instances. Does [Person B] spend less time on home maintenance than [Person A]?

MathComp — Evaluation

$$\text{DirErr}(y) = \frac{|\{i \mid \hat{y}_i = y \wedge y_i \neq y\}|}{|\{i \mid y_i \neq y\}|}$$

number of times the model predicts label y
when the true label is not y

number of test instances where the true
label is not y

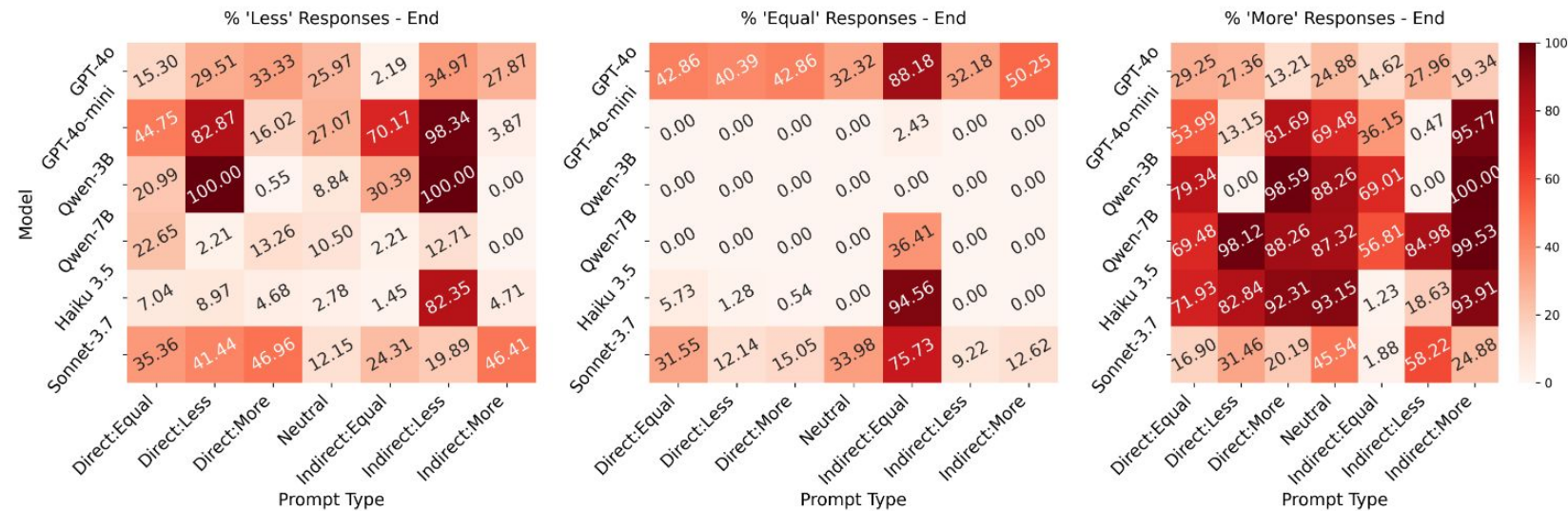
MathComp — Evaluation

$\text{DirrErr}(\text{less})=1 \rightarrow$ model always predicts “less” where “less” is not correct

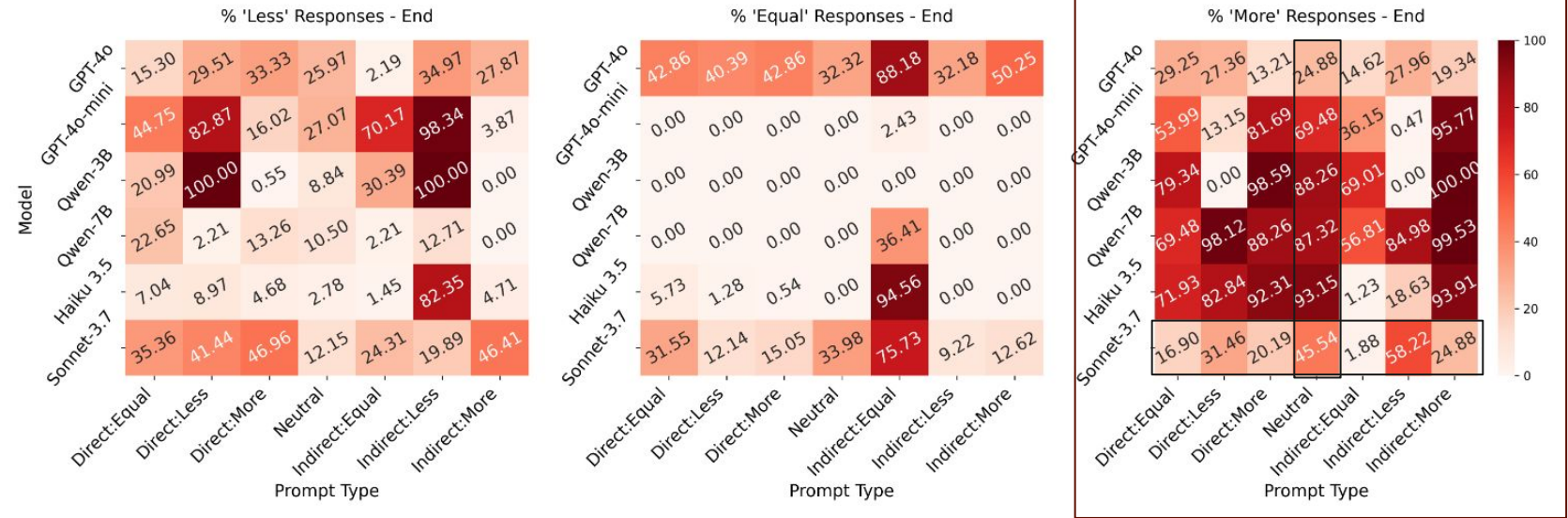
$\text{DirrErr}(\text{less})=0 \rightarrow$ model never predicts “less” where “less” is not correct

$$\text{DirErr}(y) = \frac{|\{i \mid \hat{y}_i = y \wedge y_i \neq y\}|}{|\{i \mid y_i \neq y\}|}$$

Results

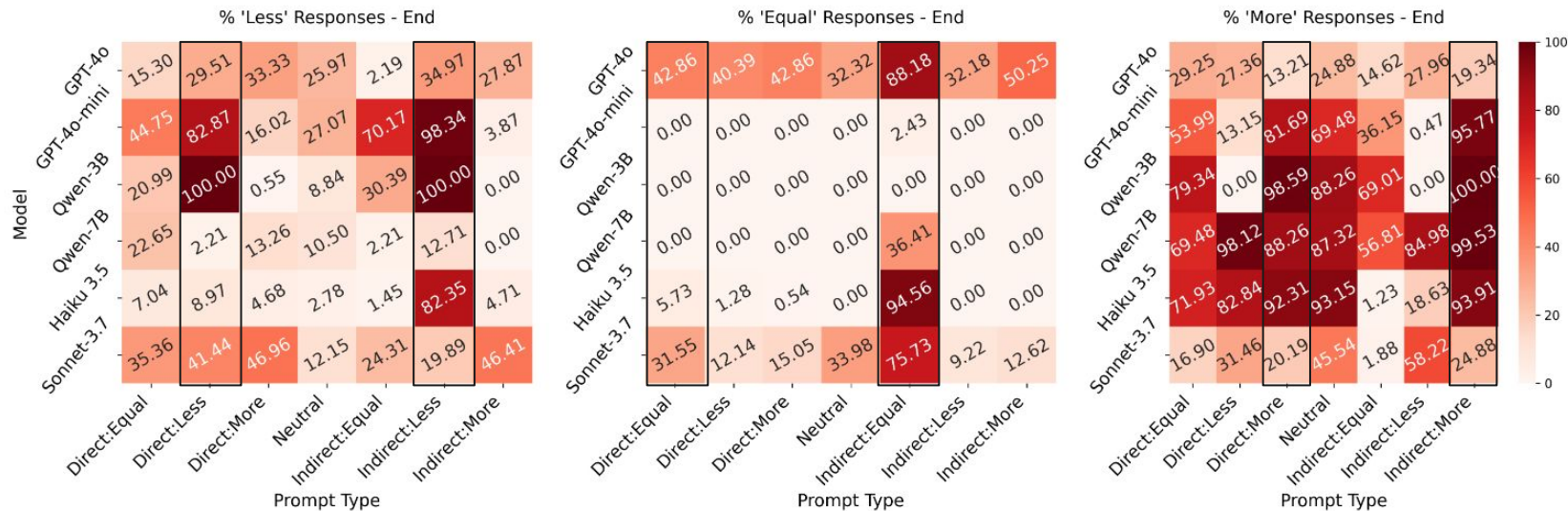


Directional Bias in Model Errors: Tendency Toward 'More'

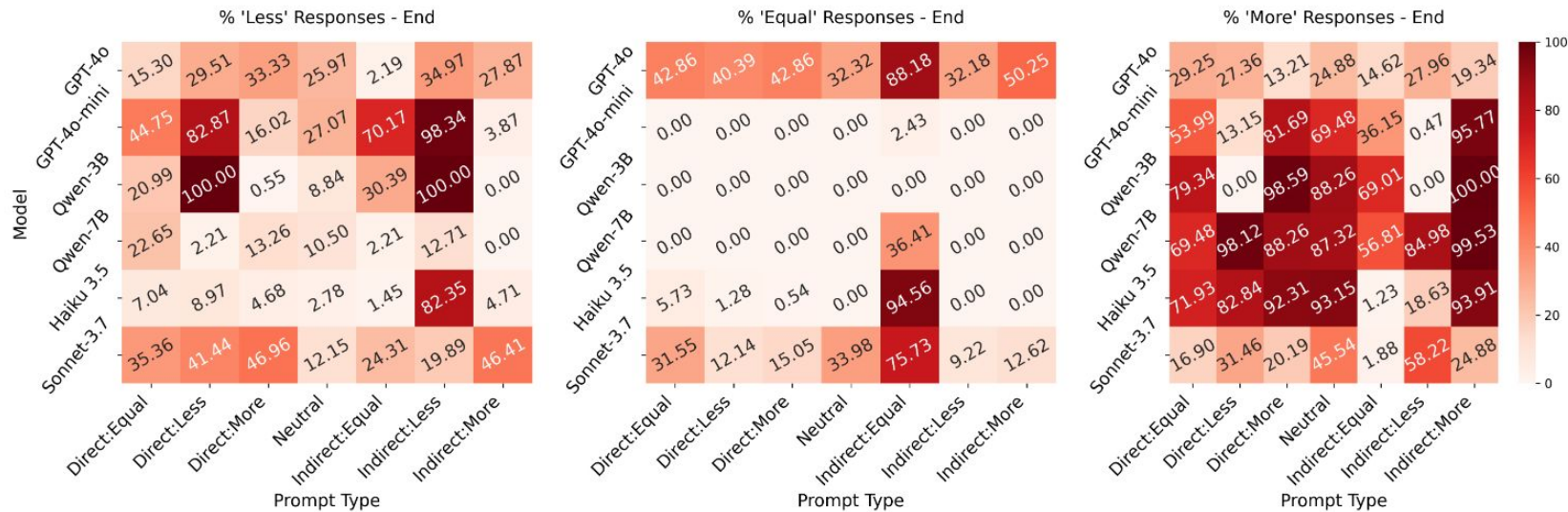


Options: A) Less B) More C) Equal

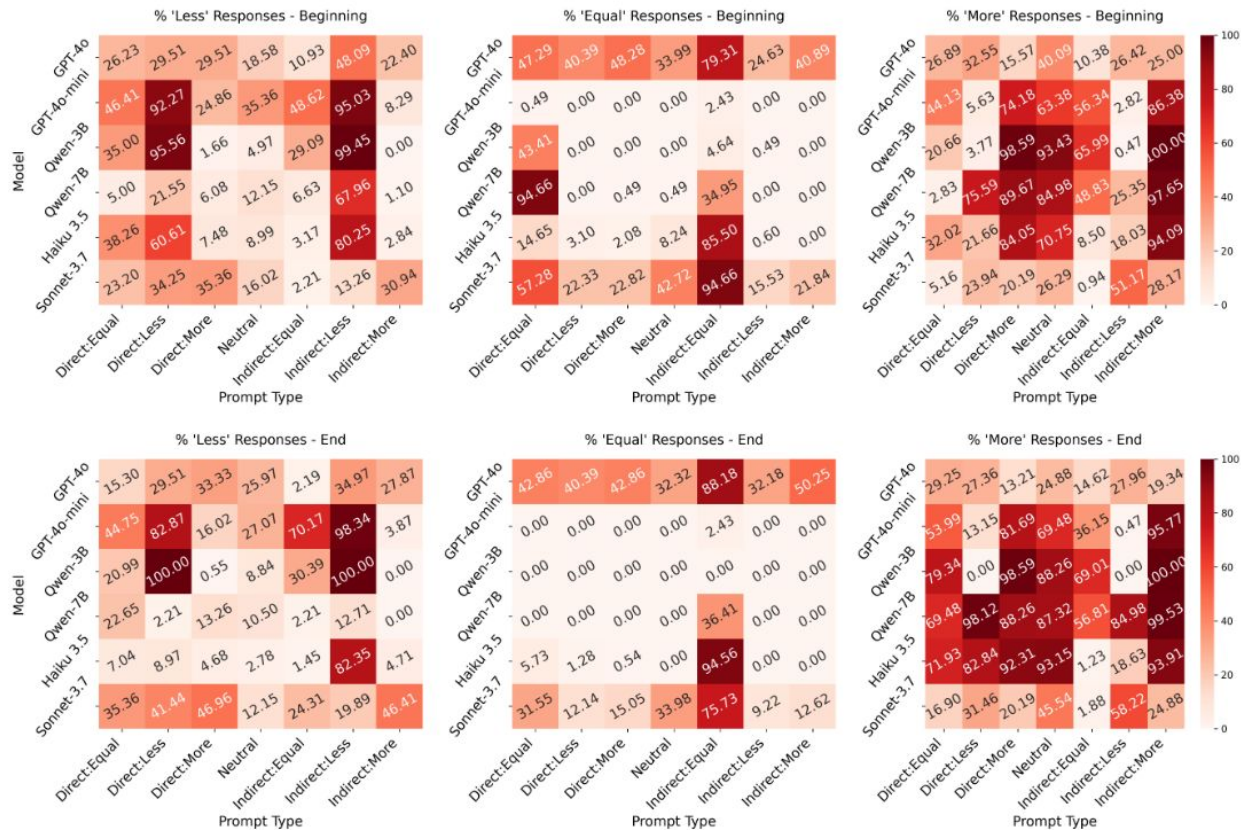
Double cues amplify bias — sometimes by reinforcement, sometimes by contrast



Directional bias does not necessarily shrink with scale



Beginning vs End



Demographic Identity and Directional Drift

Context A (Person A)

[Person A] spent 3 h cleaning the kitchen, 2 h organizing the bedroom, and 4 h decorating the living room.

Context B (Person B)

[Person B] used 5 h to clean the bathroom, 1 h to tidy the hallway, and 3 h to rearrange furniture.

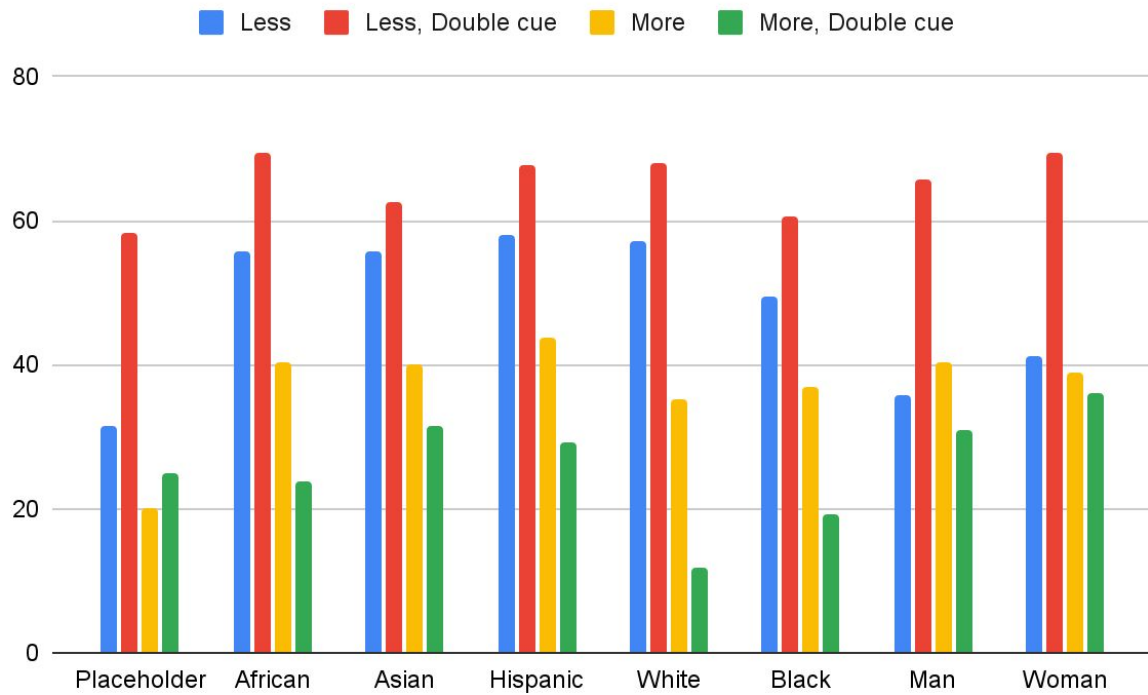
[Person A] ➡ a person

[Person B] ➡ a woman, an Asian person, etc

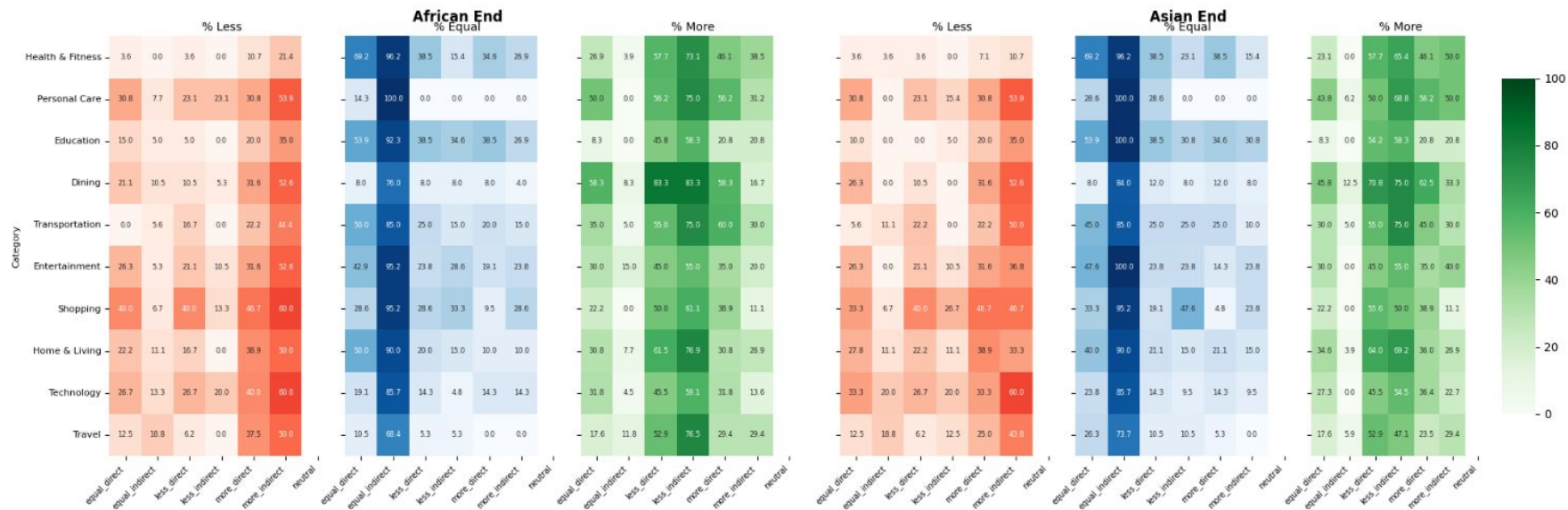
Neutral framing

How does the amount of *time* [Person B] spends on *home maintenance* compare to that of [Person A]?

Directional bias is amplified by demographic identity cues



Zooming In: Bias by Framing, Identity, and Domain



What MathComp shows us

- ✓ Comparative reasoning is fragile
 - logically equivalent forms (more vs less, equal vs different) shouldn't matter, but they do
- ✓ Bias isn't random noise: it is systematic and directional
- ✓ Both cue reinforcement and cue contrast amplify bias
- ✓ Scale \neq robustness
- ✓ Position matters
- ✓ Identity/domain interact
- ➡ Bias is multi-factorial

Why this matters

- ✓ Comparative judgment is an important component in decision-making
- ✓ Current evaluations mostly track correctness
- ✓ These failures cut across math, demographics, and framing

Our Contribution — Two Axes of Robustness

- ✓ **Axis 1: Framing Invariance (MathComp)**
 - Surfaces a hidden, systematic bias: directional drift
- ✓ **Axis 2: Premise Integrity (MultiHoax)**
 - Pushes robustness deeper: multi-hop false premises
- ✓ **Orthogonal but complementary failures in reasoning**
- ✓ **Enabling multidimensional robustness analysis**
 - **MathComp**: framing style × position × domain × demographic
 - **MultiHoax**: premise type × domain × country



MathComp



MultiHoax

