

A Good Plan is Hard to Find: Aligning LLMs with Preferences is Misaligned with What Helps Users

Nishant Balepur Matthew Shu Yoo Yeon Sung Seraphina Goldfarb-Tarrant
Shi Feng Fumeng Yang Rachel Rudinger Jordan Boyd-Graber

 @NishantBalepur  nbalepur@umd.edu



Paper



Code



GW

cohere

RLHF + Chatbot Arena rely on human preferences to make LLMs more “helpful”...

but preferences don’t capture what helps users at all!

Question	Plan A	Plan B
On the weekend, Tony will walk to the store. On weekdays, he runs to the store. When he walks, he goes 2 MPH. When he runs he goes 10 MPH. The store is 4 miles away. If he goes on Sunday, Tuesday, and Thursday, what is the average time in minutes he spends?	1. Find hours needed for walking trips by dividing distance by walking speed 2. Find hours needed for running trips by dividing distance by running speed 3. Find average time in minutes by combining one walking trip and two running trips, dividing by total trips	1. In one step, compute total minutes for Sunday's walk and Tuesday/Thursday's runs, then average them 2. Round the result from Step 1 to make it easier to report
Can you predict helpfulness?		

Can you predict helpfulness?

 Step 1: Build the *Planorama* interface

The interface shows a question about the capital of the state containing the tallest mountain in the US, followed by three steps: Step 1 (Find the tallest mountain in the US), Step 2 (Find the state that contains this mountain), and Step 3 (Find the capital of this state). It includes an "Answer (Enter)" button and a "Copy to Tool" button.

The interface shows a search for "Mount Denali" with results from "Web Page Viewer" and "Wikipedia + Calculator Search Tools". The Wikipedia page provides details about Mount McKinley, including its elevation and location.

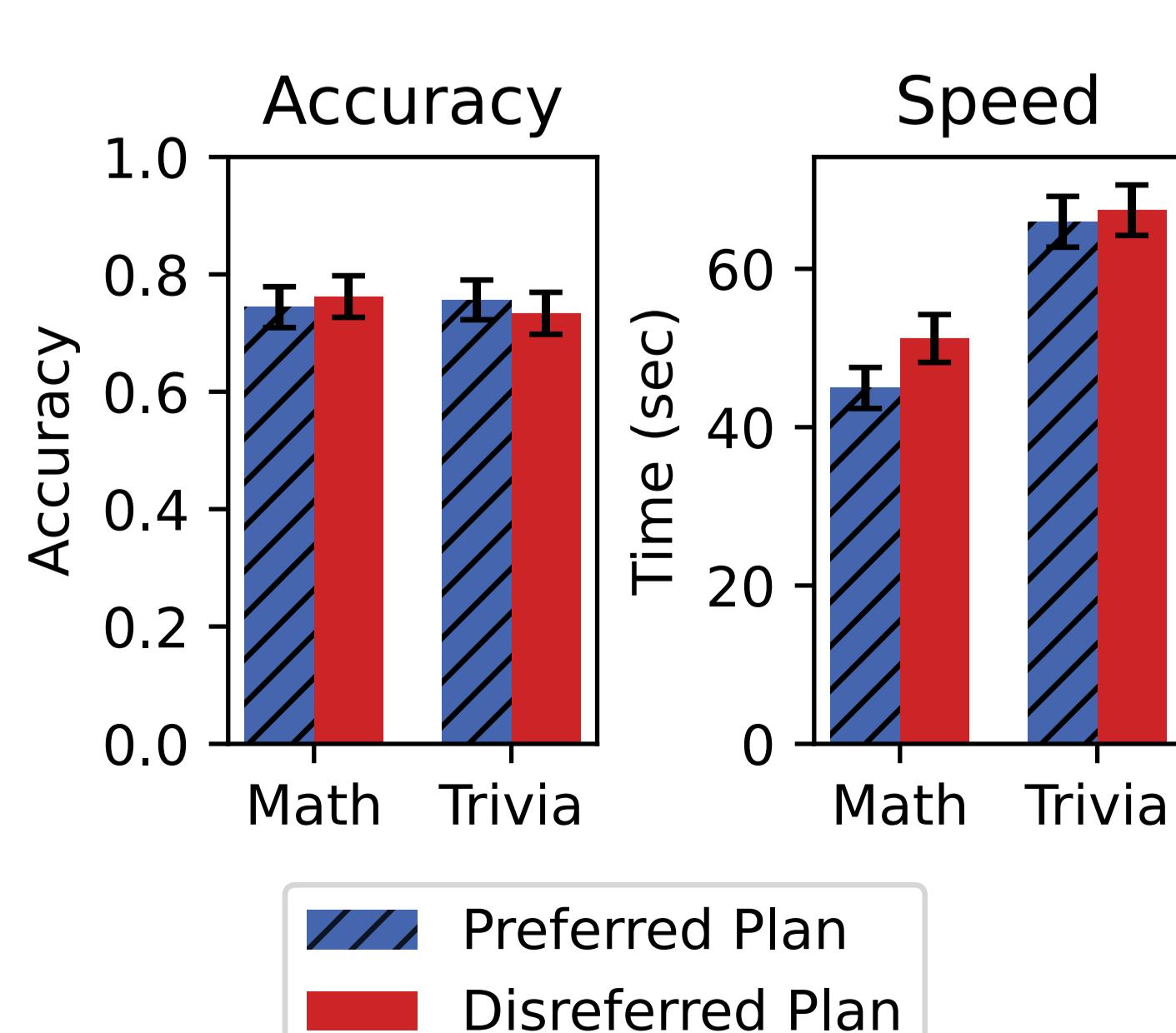
- We have 126 users:
1. Pick the plan they *think* would best help them
 2. Answer a math or trivia question w/ an LLM plan
 3. Give users access to Wikipedia + Calc. tools
 4. Record their accuracy + speed (combined via IRT) to figure out which plan *actually best helped them*

 Step 2: Show offline signals completely fail at capturing helpfulness

Proxy	Math Questions				Trivia Questions				
	User Prefer	User Helpful	GPT Prefer	GPT Helpful	User Prefer	User Helpful	GPT Prefer	GPT Helpful	
What users prefer + what helps users	User Prefer	—	52.000	67.333	55.333	—	55.667	68.667	49.000
	User Helpful	52.000	—	58.667	57.333	55.667	—	56.333	62.667
What GPT prefers + what helps GPT/ReACT	GPT Prefer	67.333	58.667	—	52.667	68.667	56.333	—	54.333
	GPT Helpful	55.333	57.333	52.667	—	49.000	62.667	54.333	—
What Reward Models score as more helpful	QRM	60.000	56.000	72.000	42.667	65.667	51.333	56.333	36.667
	GRM	53.333	54.667	66.000	38.667	64.333	51.333	57.000	40.667
	Skywork	66.667	51.333	71.333	43.333	66.333	53.333	59.000	40.000
	Nemotron	54.000	60.000	66.667	40.000	59.667	50.667	53.667	34.667
	InternLM2	57.333	57.333	70.667	41.333	61.000	52.667	51.667	38.000
	ArmoRM	56.667	56.667	68.000	39.333	59.667	52.000	53.000	42.667

Table 1: Agreement between plans user/model prefer versus plans that help users/models. *Nothing reliably predicts what helps users.*

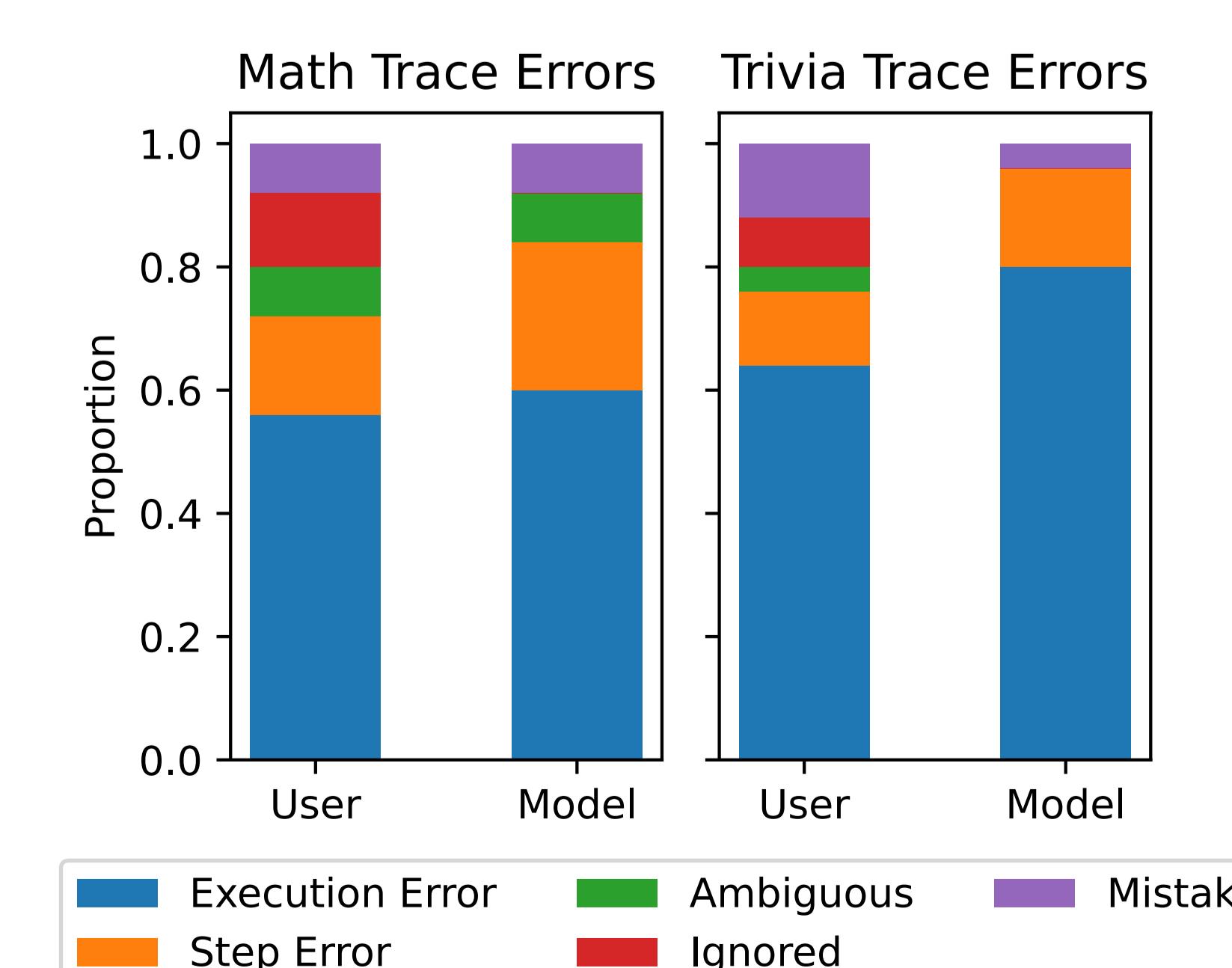
 Step 3: Figure out why this happens



Users are not much better with plans they prefer

Feature	User Prefer	User Help.	Model Prefer	Model Help.	
	Adj. R ²	0.137	0.052	0.578	0.031
Num. Steps	-0.08 (0.00)	-0.20 (0.02)	0.17 (0.00)	-0.19 (0.02)	
Num. Words	-0.04 (0.00)	-0.07 (0.02)	-0.03 (0.00)	-0.03 (0.33)	
Q/Plan Sim.	0.30 (0.00)	0.41 (0.36)	0.50 (0.00)	-0.26 (0.52)	
Diversity	-0.23 (0.29)	-0.09 (0.90)	-0.18 (0.47)	-0.97 (0.15)	
Readability	0.00 (0.52)	-0.01 (0.01)	0.00 (0.76)	-0.00 (0.86)	

Table 2: What features of plans predict preferences versus helpfulness. Users are *misled by shallow shortcuts*, which don't predict helpfulness.



Helpfulness errors are not just due to incorrectness!

Let's think beyond preferences for alignment!