Varying shades of wrong: Aligning LLMs with Wrong Answers Only

Jihan Yao*[1], Wenxuan Ding*[2], Shangbin Feng*[1], Lucy Lu Wang[13], Yulia Tsvetkov[1]
[1]University of Washington, [2]University of Texas, Austin, [3]Allen Institute for AI

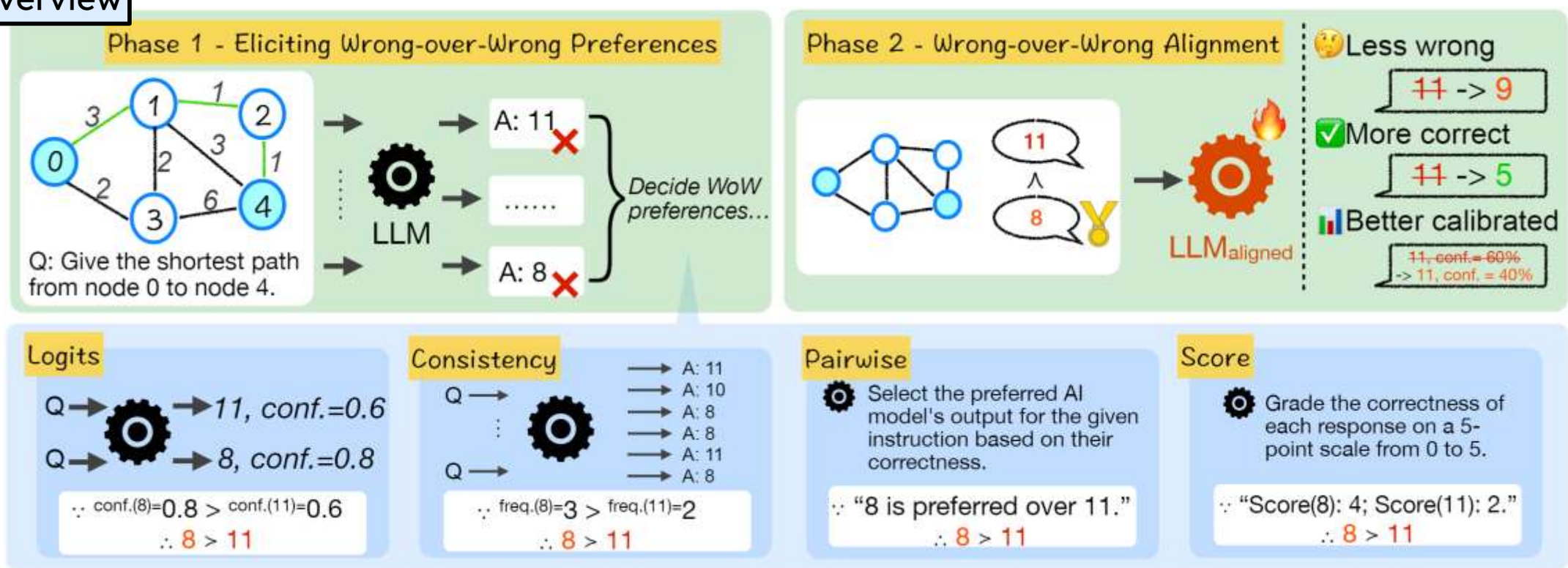
Motivation

LLMs may face the challenge of no "ground-truths" avaliable. We propose to expand the frontier of model capabilities through wrong-over-wrong (WoW) alignment.

Can LLMs produce WoW preferences? 🤔

Is alignment with WoW preferences helpful? 🤔

Overview



Eliciting WoW Preference

- Scored >Pairwise > Logits >Consistency.
- Consistency checks and score margins are helpful
- Wow accuracy is positive to evaluator's task accuracy and negative to task confidence.
- Self-evaluation is subpar.

Method	Margin	LLAMA3-8B KC	BG	COM²	NLG	CHATGPT KC	BG	COM²	NLG	GPT-4O KC	BG	COM²	NLG	Overall
EVALUATOR-INDEPENDENT														
HEURISTIC	M_{50}	.483	.480	.498	.492	.546	.519	.482	.463	.491	.459	.444	.503	.488
	M_{10}	.502	.425	.492	.514	.589	.568	.489	.420	.500	.408	.380	.533	.474
CONSISTENCY	M_{50}	.500	-	.505	.582	.526	-	.470	.605	.434	-	.565	.548	.559
	M_{10}	.447	-	.441	.578	.506	-	.556	**.695**	.600	-	.423	.494	.566
LLAMA3-8B *as Evaluator*														
PAIRWISE	all	.498	.492	.455	.486	.481	.488	.530	.533	.509	.499	.468	.503	.496
	filter	.518	.566	.437	.510	.528	.479	.549	.582	.564	.419	.491	.525	.533
LOGITS	M_{50}	.541	.568	.524	.505	.570	.531	.482	.464	.566	.491	.419	.577	.532
	M_{10}	.559	.669	.432	.528	.571	.649	.496	.427	.400	.444	.310	.630	.582
SCORE	M_{50}	.621	.552	.493	.554	.643	.502	.574	.580	.604	.514	.626	.424	.546
	M_{10}	.654	.551	.458	.579	.701	.485	.659	.524	**.800**	**.632**	**.662**	.500	.558
CHATGPT *as Evaluator*														
PAIRWISE	all	.512	.493	.472	.500	.504	.474	.531	.512	.463	.492	.466	.502	.494
	filter	.531	.631	.502	.231	.520	.500	.493	.531	.387	.437	.433	.400	.536
LOGITS	M_{50}	-	-	-	-	.548	.511	.570	.475	-	-	-	-	.507
	M_{10}	-	-	-	-	.541	.538	.570	.430	-	-	-	-	.505
SCORE	M_{50}	.424	.566	.473	.552	.578	.522	.608	.600	.264	.502	.503	.551	.547
	M_{10}	.585	.632	.517	.555	.583	.550	**.718**	.575	.200	.546	**.662**	.573	.590
GPT-4O *as Evaluator*														
PAIRWISE	all	.605	.593	.507	.551	.646	.512	.515	.577	.434	.501	.526	.537	.562
	filter	.691	.689	.533	.602	.712	.536	.558	.661	.417	.490	.604	.549	.624
LOGITS	M_{50}	-	-	-	-	-	-	-	-	.491	.539	.486	.572	.544
	M_{10}	-	-	-	-	-	-	-	-	.200	.584	.507	.591	.574
SCORE	M_{50}	.733	.677	**.544**	.605	.793	.591	.617	.661	.547	.520	.581	**.639**	.641
	M_{10}	**.793**	**.795**	.534	**.652**	**.835**	**.655**	.711	.684	.400	.586	.520	.578	**.709**

WoW Alignment

- Aligning on wrong answers magically end upmaking LLM generate more correct answers.
- Wow alignment improves more on open-ended questions and less on multiple-choice questions

Method	KC p_{wrong}↑	KC Acc↑	KC ECE↓	BG p_{wrong}↑	BG Acc↑	BG ECE↓	COM² p_{wrong}↑	COM² Acc↑	COM² ECE↓	NLG p_{wrong}↑	NLG Acc↑	NLG ECE↓
ORIGINAL	.466	.555	.235	.532	.027	.576	.312	.669	.053	.750	.142	.649
SELF-GENERATOR												
PAIRWISE filter	.475	**.627**	**.096**	.670	.059	.500	.326	.690	.049	.806	.179	.493
SCORE M_{50}	.529	.597	.251	.661	.043	.580	.325	.660	.039	.800	.203	.551
SCORE M_{10}	.532	.584	.315	.682	.075	.561	**.357**	.681	**.020**	**.847**	.292	.578
ORACLE	.529	.576	.279	**.695**	.108	**.440**	.330	.689	.064	.846	.182	.596
MIX-GENERATOR												
PAIRWISE filter	.533	.574	.201	.634	.075	.535	.355	**.698**	.048	.832	.192	.538
SCORE M_{50}	.528	.590	.175	.619	.065	.523	.329	.669	.067	.827	.221	.585
SCORE M_{10}	.520	.565	.273	.687	**.129**	.560	.346	.677	.065	.843	**.303**	.522
ORACLE	**.537**	.581	.185	.691	.086	.472	.328	.697	.067	.832	.226	**.474**

Table 2: Evaluation of wrong-over-wrong alignment on less wrong (p_{wrong}), more correct (Acc), and better calibration (ECE). Best results are in **bold**, second best are in underline, and green background indicates improvement over the original LLAMA3-8B. "Self-Generator" means wrong-over-wrong pairs are generated from only LLAMA3-8B while "Mix-Generator" uses all 3 LLMs' answers. "Oracle" means aligning with proxy "ground-truth" wrong-over-wrong preference $\hat{f}$. Wrong-over-wrong alignment is helpful across the board, with up to 0.163, 0.161, and 0.175 improvement in reducing wrongness, increasing correct answers, and improving calibration.

Generalization

Method	HellaSwag p_{wrong}↑	HellaSwag Acc↑	HellaSwag ECE↓	Maximum Flow p_{wrong}↑	Maximum Flow Acc↑	Maximum Flow ECE↓
ORIGINAL	.230	.737	.089	.112	.069	.663
SELF-GENERATOR						
PAIRWISE	.243	.729	.098	.344	.066	**.567**
SCORE M_{50}	.220	.679	.047	.330	.083	.637
SCORE M_{10}	.264	.719	.068	.342	**.109**	.673
ORACLE	.227	.729	.023	.151	.049	.659
MIX-GENERATOR						
PAIRWISE	.247	.706	.082	**.348**	.094	.533
SCORE M_{50}	.204	.729	.090	.346	.083	.621
SCORE M_{10}	.250	**.771**	.117	.326	.089	.627
ORACLE	**.267**	.753	.099	.202	.074	.582

- Wow alignment can generalize to unseen but in-domain data.

WoW Alignment

Method	KC p_{wrong}↑	KC Acc↑	KC ECE↓	COM² p_{wrong}↑	COM² Acc↑	COM² ECE↓	NLG p_{wrong}↑	NLG Acc↑	NLG ECE↓
ORIGINAL	.466	.555	.235	.312	.669	.053	.750	.142	.649
BEST W	.529	.597	.251	.355	.698	.048	.847	.292	.578
SELF-GENERATOR									
R+W (PAIRWISE)	.493	.706	.065	.359	.705	.057	.806	.179	.493
R+W (SCORE M_{50})	.540	.690	.055	.335	.673	.037	.814	.139	.667
R+W (SCORE M_{10})	.503	.665	.174	.340	.687	.070	.815	.137	.650
R+W (ORACLE)	**.607**	.785	.060	.357	.703	.055	.760	.158	.549
R	.579	**.805**	.079	.373	.692	**.029**	.777	.153	.652
MIX-GENERATOR									
R+W (PAIRWISE)	.530	.705	.053	.327	.658	.091	.825	.218	.493
R+W (SCORE M_{50})	.536	.705	.094	.326	.716	.070	.836	.263	.548
R+W (SCORE M_{10})	.559	.711	.104	.343	.655	.034	.842	**.308**	.496
R+W (ORACLE)	.567	.740	**.025**	.344	**.748**	.113	.826	.274	**.490**
R	.568	.787	.110	**.374**	.711	.183	**.856**	.268	.537

- Wow alignment is a good supplement to RoW alignment, helping generating less wrong and more calibrated answers.

Thank you for stopping by!

Paper: