Arguments that Alter Minds: LLM Rationales Sway Human (and LLM) Notions of Plausibility

Shramay Palta, Peter Rankel,
Sarah Wiegreffe and Rachel Rudinger

Problem Setting

- We use MCQ benchmarks for evaluating commonsense reasoning.
- Commonsense reasoning answer choices lie on a continuum of plausibility.

- An answer choice can be subject to arguments *for* or *against* its plausibility.
- Such arguments *do not add* any new information.
- Simply highlight circumstances, which if true, would impact the answer's plausibility.

Rationale Generation

- Use 4 different models to generate PRO and CON Rationales.
- Preference study with 4 annotators who are tasked to choose the best rationale.
- GPT-4o received most votes!

Methodology

- Two datasets: Social IQa and CommonsenseQA.
- Use NO Rationale ratings collected by Palta et al. 2024.
- Ask humans and LLMs to rate plausibility of gold-label and distractor answer choices after adding PRO, CON or PRO+CON Rationales.
- 3000 human and 13,600 LLM ratings collected!

How do Rationales affect Plausibility Judgments?

Dataset	Agent	Pro Rationale			Con Rationale			Pro+Con Rationales		
		Overall	Gold Label	Distractor	Overall	Gold Label	Distractor	Overall	Gold Label	Distractor
SIQA	Human	3.33(+0.1)	3.84(−0.26)	2.81(+0.47)	2.28(−0.94)	2.72(−1.39)	1.85(−0.49)	2.87(−0.35)	3.36(−0.75)	2.39(+0.05)
	OpenAI	3.7(+0.57)	4.25(+0.26)	3.14(+0.88)	1.9(−1.22)	2.28(−1.7)	1.52(−0.74)	2.9(−0.22)	3.38(−0.6)	2.42(+0.16)
	Non-OpenAI	3.6(+0.62)	3.94(+0.42)	3.26(+0.82)	2.21(−0.77)	2.42(−1.09)	1.99(−0.45)	2.93(−0.05)	3.19(−0.32)	2.66(+0.22)
CQA	Human	3.39(0.0)	3.91(−0.44)	2.86(+0.45)	2.62(−0.76)	3.28(−1.08)	1.96(−0.45)	2.97(−0.41)	3.7(−0.65)	2.24(−0.17)
	OpenAI	3.91(+0.64)	4.55(+0.25)	3.27(+1.03)	2.02(−1.25)	2.56(−1.74)	1.48(−0.76)	3.12(−0.15)	3.76(−0.54)	2.49(+0.25)
	Non-OpenAI	3.63(+0.61)	3.95(+0.31)	3.32(+0.92)	2.18(−0.84)	2.43(−1.21)	1.94(−0.47)	2.99(−0.03)	3.28(−0.36)	2.7(+0.3)

Table 1: Plausibility ratings after different types of rationales were shown to humans and LLMs.

Overall

- PRO Rationales lead to an increase in ratings.
- CON Rationales lead to a decrease in ratings.
- PRO+CON Rationales illicit a mixed response, with ratings likely to drop or stay unchanged.

🔍 Taking a closer look...

- PRO Rationales have a bimodal effect on human ratings. Distractor ratings *rise*, but gold label ratings *fall*!
- PRO+CON Rationales push ratings to the center of the scale, for both humans and LLMs.

What caused the Plausibility Judgments to change?

		Feature							
		NO Rationale Rating		PRO Rationale		CON Rationale		PRO+CON Rationale	
Dataset	Agent	Gold Label	Distractor	Gold Label	Distractor	Gold Label	Distractor	Gold Label	Distractor
SIQA	Human	**−0.4581**	**−0.5495**	**0.2417**	0.1082	**−0.8823**	**−0.8525**	**−0.2423**	**−0.3125**
	OpenAI	**−0.4503**	**−0.5086**	**0.7083**	**0.5023**	**−1.2592**	**−1.1177**	**−0.1592**	**−0.2127**
	Non-OpenAI	**−0.4770**	**−0.6491**	**0.6659**	**0.4543**	**−0.8452**	**−0.8102**	−0.0785	**−0.1391**
CQA	Human	**−0.5092**	**−0.4983**	0.2445	**0.1590**	**−0.3875**	**−0.7450**	0.0405	**−0.4650**
	OpenAI	**−0.2340**	**−0.4883**	**0.5517**	**0.6614**	**−1.4408**	**−1.1286**	−0.2383	**−0.1236**
	Non-OpenAI	**−0.5632**	**−0.6236**	**0.6693**	**0.5428**	**−0.8507**	**−0.8394**	−0.0040	−0.0750

Table 2: OLS regression coefficients for different features to understand why the judgments changed after introducing rationales.

- OLS Regression to understand fluctuations in ratings.
- Dependent Variable: Change (Δ) in plausibility rating after adding a rationale.
- Features: Rationale Type, and NO Rationale rating.

- NO Rationale rating exhibits a strong anchoring effect! More pronounced for distractor answer choices.
- CON Rationales exert a stronger effect as compared to PRO Rationales.

Key Takeaways

- LLMs rationales can be very persuasive, even for incorrect answer choices.
- Humans and LLMs respond differently to the addition of these rationales.