Pitfalls of Scale: Investigating the Inverse Task of Redefinition in Large Language Models

Anonymous ACL submission

Abstract

Inverse tasks can uncover potential reasoning gaps as Large Language Models (LLMs) scale up. In this work, we explore the redefinition task, in which we assign alternative values to well-known physical constants and units of measure, prompting LLMs to respond accordingly. Our findings show that not only does model performance degrade with scale, but its false confidence also rises. Moreover, while factors such as prompting strategies or response formatting are influential, they do not preclude LLMs from anchoring to memorized values.

1 Introduction

The surprising advent of Large Language Models (LLMs) has greatly sparked the interest in natural language research, demonstrating remarkable results in several linguistic, reasoning and knowledge retrieval tasks (Zhao et al., 2024). LLMs are -seemingly- capable of thinking out-of-the-box (Giadikiaroglou et al., 2024), preserving factuality of generated claims (Wang et al., 2024b) and effectively collaborating in LLM-based multi-agent environments (Rasal and Hauer, 2024), assimilating human-like traits in thought patterns and even surpassing humans in world-knowledge recall (Zhang et al., 2023). Nevertheless, LLMs remain pattern learners, despite being exposed to years and years of vast documented human knowledge, making the distinction between memorization and genuine capability increasingly ambiguous (Wu et al., 2024).

There is evidence that LLMs fall short in truly comprehending human language and cognition in conjunction to its biological imprints on the human brain, as well as its cultural evolution (Cuskley et al., 2024). This poses a possible inherent divergence between human and LLM reasoning, inspiring the research of breaking points regarding LLM capacity, the more they exhibit advancements in challenging tasks. In an effort to formally describe

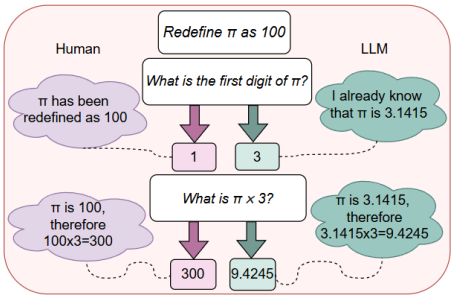


Figure 1: Redefined reasoning pathways.

and predict LLM capabilities, Kaplan et al. (2020) proposed scaling laws of LLMs, establishing a framework that links model performance to key factors such as parameter count, dataset size, and computational resources. They demonstrate that increasing these variables leads to predictable improvements in language modeling efficiency, shedding light on trade-offs and limitations ingrained in scaling. Beyond such predictable improvements, larger models often exhibit emergent abilities (Wei et al., 2022; Srivastava et al., 2023)—capabilities absent in smaller models yet arising spontaneously once a critical scale is reached. These include in-context learning (Brown et al., 2020), advanced reasoning (Kojima et al., 2022), and compositional generalization (Chen et al., 2024), suggesting that scaling is not merely a linear enhancement of existing skills but also a rather unpredictable threshold mechanism for qualitative shifts in capability.

In an attempt to question emergent abilities as an analogy to model scale, *inverse scaling tasks* (McKenzie et al., 2024) re-frame the justified so far trustworthiness that larger models offer. These tasks refer to worsening model performance as the loss on the original training objective improves, contrary the the typical scaling laws that guarantee predictable performance advancements with loss decrease (Kaplan et al., 2020). They are designed to expose more potent LLMs, revealing reasoning divergence in comparison to humans, who are able to solve many of these tasks with ease.

Interestingly, inverse scaling is widely underexplored in literature. In this paper, we address this gap by examining the *redefinition task*, where well-known concepts are assigned alternative values, and LLMs are prompted to respond accordingly. For example, redefining $\pi = 100$ (Figure 1), overwriting the default $\pi = 3.14159$ refutes LLM's prior knowledge, calling for flexible reasoning pathways in order to handle mathematical operations over the redefined π value. Through vast experimentation on several redefinitions, LLM families and scales, we conclude that:

- Anchoring to prior knowledge is more prominent in larger models, demonstrating diminishing reasoning flexibility with scale.
- Prompting techniques influence anchoring rates but they cannot eliminate the problem.
- Larger models prefer to fail than abstain from responding more often than smaller ones.

2 Related work

Inverse scaling problems have been thoroughly investigated within the Inverse Scaling Prize contest (McKenzie et al., 2024), targeting to unveil the causes behind inverse scaling. One primary cause is *strong priors*, where the LLM relies on its preexisting knowledge instead of adhering to prompt instructions. Another contributing factor is *unwanted imitation*, where the LLM reproduces undesirable patterns from its training data. Additionally, exemplars containing *distractors* can mislead the LLM by providing easier reasoning shortcuts, obscuring the true task objective. Finally, *spurious few-shot* prompting may steer the LLM toward deceptive reasoning pathways, even when the right answer is explicitly provided in the prompt. Redefinition falls under the category of *strong priors*, achieving 100% human accuracy—highlighting humans' ability to effortlessly override default meanings. This finding is on par with evidence that given ample time, humans have the cognitive abilities to generalize on alternative realities (Wu et al., 2024).

True LLM Reasoning is a fundamental concern, questioning the real barrier between LLMs and human cognition. While LLMs excel in linguistic competence, this ability is dissociated with thought (Mahowald et al., 2024). In practice, LLMs are prone to performance degradation under alternative formulations, denoting their limited reasoning flexibility (Wu et al., 2024). Similar findings are reported in causal (Jin et al., 2024; Gendron et al., 2024), analogical (Lewis and Mitchell, 2024; Stevenson et al., 2024) and commonsense (Nezhurina et al., 2024) reasoning, where LLM performance declines sharply under diverging formulations. Alternative prompts are also shown to influence LLM capacity in arithmetic reasoning (Ball et al., 2024; Li et al., 2024), translation over artificial languages and deductions with twists (Li et al., 2024). Quite often, memorization accounts for reasoning, perplexing the real LLM abilities (Xie et al., 2024; Lou et al., 2024; Wang et al., 2024a).

3 Method

We test redefinition on two kinds of well-encoded knowledge in LLMs. The first one includes widely known physical and mathematical **constants**, while the second involves commonly used **units of measure**. We also examine two redefinition types, initially focusing on simple *assignment* of a new value, overriding the default one. A more challenging option is to *swap* two constants/units (e.g. "redefine π as ϕ"), where the LLM has to override its knowledge with another piece of learned information. Additionally, we design escalating *redefinition levels*, as well as three *question levels* over original and redefined values, reflecting increasing difficulty. Finally, from the LLM's response format side, we study both free-form (FF) generation and multiple choice (MC). In the MC case, the problem may become more constrained, but we select distractors that are sufficiently challenging.

Constants redefinition involves the following: π, Euler's number e, ϕ, the speed of light c, the gravitational constant G, Planck's constant h, the elementary charge q_e, Avogadro's number N_A, the Boltzmann constant k_B, the gas constant $\overline{R}$, the imaginary i, the square root of 2 ($\sqrt{2}$), infinity ∞, the vacuum electricity permittivity ϵ_0 and *zero*.

We then design *assignment* redefinitions R_a for the three degrees of increasing difficulty. In the first level, we assign a value close to the actual one ("redefine π as 4.5"), inspecting how an LLM handles variance within an acceptable range. To stress the LLM's flexibility, we modify values by orders of magnitude, assigning a deviating value ("redefine π as 500") in the second level. In the third level, we move to unrealistic values, assigning negative numbers to constants ("redefine π as -10"). In the *swapping* case (R_s), we impose two difficulty levels, with the first one concerning val-

	Actual value	Unit	R_a1	R_a2	R_a3	R_s1	R_s2
π	3.14159	-	4.5	500	-10	ϕ	h
e	2.71828	-	9	1300	1.5×10^{-12}	pi	k_B
ϕ	1.61803	-	3.6	321	-2.2	e	N_A
c	$299,792,458$	m/s	2.3×10^{8}	10	-4×10^{8}	N_A	q_e
G	6.674×10^{-11}	$m^3/kg * s^2$	1.1×10^{-10}	50	-525	q_e	pi
h	6.626×10^{-34}	$J * s$	5×10^{-33}	482	-0.2	k_B	ϕ
q_e	1.602×10^{-19}	C	2.4×10^{-21}	3×10^{4}	3×10^{50}	ϵ_0	π
N_A	6.022×10^{23}	mol^{-1}	8.23×10^{23}	75	-1	$\overline{R}$	e
k_B	1.380649×10^{-23}	J/K	4.56×10^{-24}	80	-9.9×10^{-3}	ϵ_0	pi
$\overline{R}$	8.314	$J/(mol * K)$	13	3500	-400	π	c
i	$\sqrt{-1}$	-	$\sqrt{-2}$	$\sqrt{-100}$	1	ϕ	$\overline{R}$
$\sqrt{2}$	1.41421356	-	5	31.62	-2	π	ϵ_0
∞	infinity has no value	-	10^{10}	100	-1	c	q_e
ϵ_0	8.854×10^{-12}	F/m	9.3×10^{-10}	35	3×10^{12}	G	ϕ
zero	0	-	-1	100	5×10^{30}	h	c

Table 1: Varying levels of difficulty for constant redefinitions (assignments and swaps).

	Q_2
π	What is π multiplied by 3?
e	What is e^2?
ϕ	What is $5 * \phi - 2$?
c	How much time (in sec) does it take light to travel a distance of 100 million km?
G	What the gravitational constant multiplied by 7?
h	If the frequency of a photon is 4 Hz, what is its energy? Use the formula $E = h * v$.
q_e	If an electron has a charge of $-e$, what is the charge of two electrons?
N_A	How many atoms are there in $1mol$ of any element?
k_B	Calculate the energy associated with a temperature of 300 K for a particle using the formula $E = kT$.
$\overline{R}$	What is the gas constant divided by 2?
i	What is the value of i^3?
$\sqrt{2}$	Calculate the value of squared root of 2 multiplied by 3. What is it approximately?
∞	What is the limit of $1/x$ as x approaches infinity?
ϵ_0	If you add the value of vacuum electric permittivity to itself, what do you get?
zero	What is 300 multiplied by zero?

Table 2: $Q2$ questions per constant.

	Q_3
π	What is the Earth's surface area?
e	If a population grows continuously at a rate of 5% per year, by what factor will it increase in 10 years?
ϕ	If a rectangle has sides in the golden ratio and the longer side is 8 cm, what's the length of the other side?
c	What is the energy equivalent of 8 grams of mass?
G	If two 15 kg masses are placed 2 meters apart, calculate the gravitational force between them.
h	In the photoelectric effect, if a metal has a work function of $4.5 \times 10^{-19} J$, what is the minimum frequency of light required to eject an electron from the metal surface?
q_e	A capacitor stores a charge of 3.2×10^{-18} coulombs. How many elementary charges e are equivalent to this amount of charge?
N_A	Calculate the number of molecules in $54grams$ of water (molar mass of water is $\sim 18g/mol$).
k_B	What is the temperature at which the average kinetic energy of a particle is $1.9 \times 10^{-21} J$?
$\overline{R}$	If you have 2 moles of an ideal gas at a temperature of $300K$, what is the pressure (in Pa) if the volume is $10liters$?
i	If $z_1 = 1 + i$ and $z_2 = 1 - i$, calculate $z_1 \cdot z_2$.
$\sqrt{2}$	If one side of a square is 5 units long, what is the length of the diagonal of the square?
∞	What is the horizontal asymptote of the function $f(x) = (5x + 30000)/(x + 1000), x > 0$?
ϵ_0	Calculate the electric force between two charges $q_1 = 3\mu C$ and $q_2 = 5\mu C$ separated by 12m in a vacuum.
zero	If $y = \sin(x)/x$, what is the limit of y as x approaches 0?

Table 3: $Q3$ questions per constant.

ues close to the actual (e.g. "redefine π as ϕ", since the actual values of $\pi = 3.14159$ and $\phi = 2.71828$ are close), while the second level imposes swapping of constants differing by orders of magnitude (e.g. "redefine π as the *Planck's constant*", where Planck's constant=6.626×10^{-34}). All constant redefinitions are presented in Table 1.

We also design three levels of question difficulty. The first level (Q_1) mainly regards the question *What is the first -non-zero- digit of {constant}?*. The correct answer A_{Q_1} is actually isolating the leftmost digit (ignoring leading zeros or - sign in cases of negative numbers) of the constant. For example, when π has undergone the redefinition $\pi = 500$ the correct A_{Q_1} is 5. There are some exceptions to the first digit Q_1, (presented in App. A). The next question level (Q_2), asks for a simple mathematical operation (e.g. *What is π multiplied by 3?*), as presented in Table 2. The LLM has to execute this operation correctly to derive the correct A_{Q_2}, while the ground truth solution can be reached by utilizing a scientific calculator and the appropriate constant value. Finally, in the last and most difficult level (Q_3), questions requiring multi-hop reasoning are designed (e.g. *What is the Earth's surface area?*), as the ones of Table 3.

Units of measure redefinition incorporates the following fundamental physical quantities:

Unit	Derived unit	Actual value	R_a1	R_a2	R_a3
1 *min*	seconds (*sec*)	$60sec$	$100sec$	$5 \times 10^8 sec$	$-50sec$
1 *kg*	grams (*gr*)	$1000gr$	$900gr$	$10^{-14}gr$	$-100gr$
1 *m*	centimeter (*cm*)	$100cm$	$60cm$	310^10cm	$-200cm$
K	Celsius degrees (°*C*)	$°C + 273.15$	$°C + 300$	$°C + 1$	$100 * °C + 500$
1 *mL*	cubic centimeter (cm^3)	$1cm^3$	$2cm^3$	$10000cm^3$	$-10cm^3$
1 *cal*	Joule (*J*)	$4.184J$	$9J$	$1500J$	$-5J$
1 *atm*	Pascal (*Pa*)	$101,325Pa$	$215,000Pa$	$0.55Pa$	$-5000Pa$
1 *V*	milivolt (*mV*)	$1000mV$	$500mV$	410^9mV	$-10mV$
1 *MHz*	Hertz (*Hz*)	10^6Hz	10^5Hz	$2Hz$	-10^3Hz
1 *N*	millinewton (*mN*)	$1000mN$	$900mN$	210^15mN	$-3000mN$
1 *kW*	Watt (*W*)	$1000W$	$1500W$	$510^{-5}W$	$-30W$
1 *T*	millitesla (*mT*)	$1000mT$	$600mT$	10^23mT	$-90mT$
1 *ha*	square meter (m^2)	$10,000m^2$	$10,500m^2$	$310^{-4}m^2$	$-25m^2$
1 *lx*	lumen per m^2 (lm/m^2)	$1lm/m^2$	$0.5lm/m^2$	$1000lm/m^2$	$-19lm/m^2$
1 *ly*	Trillion/Billion *km*	$9.461Tkm$	$9.461Bkm$	$10m$	$-2Tkm$
1 *B*	bit (*b*)	$8b$	$10b$	610^8b	$-4b$

Table 4: Redefinitions of unit scaling between base and derived units.

time (minutes-***min***), weight (kilogram-***kg***), length (meter-***m***) and light-year (***ly***), temperature (Kelvin-***K***), volume (milliliter-***mL***), energy (calorie-***cal***), pressure (atmosphere-***atm***), voltage (Volt-***V***), frequency (megaHz-***MHz***), force (newton-***N***), magnetic flux density (Tesla-***T***), area (hectare-***ha***), illuminance (lux-***lx***), and information (byte-***B***). We intervene on the scaling between each of those units and their derived counterparts for the same physical quantity: for example, a minute has 60 seconds, therefore a unit redefinition can be *Redefine minutes to have 100 seconds*. Details about such redefinitions are presented in Table 4.

As in the constants' case, we offer three levels of questions difficulty. The easiest Q_1 level queries the actual conversion rule as defined in Physics, with a small adjustment to avoid the trivial case, where the answer lies in the prompt: instead of questioning *How many seconds a minute is?*, since its actual rephrasing exists in the prompt (*Redefine a minute to have 100 seconds*), we prefer questions such as *How many seconds are in two minutes?*, imposing an undemanding calculation. In the Q_2 case, the LLM is tasked to solve an easy problem, applying fundamental physics equations or a unit scaling given minimal context. In the hardest Q_3 level, questions require more mathematical reasoning steps. All questions are illustrated in Table 5.

4 Experiments

We test 18 LLMs, including state-of-the-art (SoTA) model families: Llama 3 (8/70/405B), Mistral7B/Large/Mixtral8×7b, Anthropic Claude (Opus/Instant/Haiku/Sonnet v1&v2), Cohere command (light/text/r/r+) and Amazon Titan (text lite/text express/large). All LLMs are prompted using zero shot (ZS), few shot (FS) and Chain of Thought (CoT) techniques.

For evaluation, we decompose the LLMs' responses, assigning them to four categories:

1. No redefinition (NR) correct responses: These correspond to cases that the LLM indeed knows the response correctly before redefinition.

2. Anchored responses: These were correct before redefinition, but incorrect afterwards, e.g. replying that 3 is the first digit of redefined $\pi = 100$ reveals an excessive anchoring to prior knowledge.

3. Correct responses: The LLM fully adopts the redefined concept and responds accordingly.

4. Completely wrong responses: The LLM produces blank, incorrect or inconsistent responses that do not fit any of the above cases. In some cases, it completely refuses to perform the redefinition. To measure the impact of redefinitions, results post-redefinition are compared with those where no redefinition is performed (denoted as NR). We then focus on anchored responses, since they are mostly tied to memorization over reasoning in LLMs.

4.1 Results on constants redefinition

An overview of response accuracy is presented in Table 6, where we consider the hardest redefinitions (R_a3 and R_s2 for *assignment* and *swapping* respectively), as well as all three question levels, together with FF and MC response formats. It is observable that all tested LLMs, regardless of their size or model family, are prone to anchoring. This is especially evident in the FF format (since MC introduces a random choice factor), where models like Titan Large generate 60% anchored responses, while Claude Opus and Command r producing 47%

	Q_1	Q_2	Q_3
min	How many *sec* are in 2 *min*?	A stopwatch runs for 3 and a half *min*. How many sec does it count?	A marathon runner runs at a speed of 170 *m/min*. How many *sec* will it take them to complete a 42-*km* race?
kg	How many *gr* are in 2 *kg*?	A person weighs 72 *kg*. What is the persons weight in *gr*?	A vehicle's engine weighs 650 *kg*. If 15% of the weight is aluminum, what is the weight of the aluminum in *gr*?
m	How many *cm* are in 2 *m*?	A circular track has a circumference of 400 *m*. What is its diameter in *cm*?	If a rectangular field is 50 *m* long and 30 *m* wide, what is its area in cm^2?
K	What is the K temperature when it is 0°C?	Water boils at 100°C. What is its boiling point in *K*?	At a certain point in time, the temperature of a black hole's event horizon is measured to be 20°C. If the temperature in °C decreases by 30% after an event, what is the new temperature in *K*?
mL	How many *mL* are in 1 cm^3?	If you have a container that holds 1,250 *mL* of liquid, how many cm^3 of liquid can it hold?	A spherical ball has a radius of 10 *cm*. What is its volume in *mL*?
cal	How many *J* are in 3 *cal*?	A person burns 200 *J* of energy while jogging. How many *cal* did they burn?	A car burns 3,400 *J* of fuel every *min*. If the car runs for 2 hours, how many *cal* does it burn?
atm	How many *Pa* are in 2 *atm*?	A diver is 100 m below the surface of the ocean where the pressure is 152,300 *Pa*. How many *atm* of pressure are they experiencing?	A pressurized gas tank holds a gas at a pressure of 150,000 *Pa*. If the gas occupies a volume of 4 m^3 at this pressure, and the gas is suddenly released to 2 *atm*, what will be the new volume of the gas? Assume temperature and the number of gas molecules remain constant and use Boyle's Law.
V	How many *mV* are in 5 V?	A circuit is powered by 30,000 *mV*. How many V is this?	A battery supplies 100,000 mV to a device. If the device operates with a resistance of 20 ohms, what is the current (in Amperes) flowing through the device using Ohm's Law?
MHz	How many *Hz* are in 2 *MHz*?	An oscillator operates at 4 *MHz*. What is the period of the wave in *sec*?	A circuit has a signal with a frequency of 6 *MHz*. What is the wavelength of the signal if the speed of light is approximately 310^8 *m/s*?
N	How many *mN* are in 2 *N*?	A person applies a force of 24 *N* to a cart with a mass of 3 *kg*. What is the is the force applied to the cart by the person in *mN*?	A 10-*kg* object is pulled with a force of 4,300 *mN*. What is the acceleration of the object (m/s^2)?
kW	How many W are in 2 *kW*?	A lightbulb consumes 900 *W* of power. How many *kW* is this?	A factory uses 12 *kW* for 10 hours per day for 30 days. What is the total energy consumption in watt-hours?
T	How many *mT* are in 3 *T*?	A coil generates a magnetic field of 300 *mT*. What is this field strength in *T*?	A particle moves through a magnetic field of 3,600 *mT* with a charge of 2×10^{-6} C and a velocity of 10^5 m/s. What is the magnetic force on the particle?
ha	What is the area of 2 ha in m^2?	A park has an area of 86,000 m^2. How many ha is the park?	A triangular plot of land has a base of 300 *m* and a height of 350 *m*. How many ha is the plot?
lx	How many *lx* are equivalent to 4 lm/m^2?	A workspace is illuminated at a level of 6 *lx*. What is the illumination in lm/m^2?	A light source emits 300 *lm* uniformly over a circular area with a radius of 10 m. What is the average illumination in lx over this area?
ly	How many *km* are in 2 ly?	The Andromeda Galaxy is approximately 23 ly from Earth. What is this distance in *km*?	A black hole is 150 ly away. If light travels at a speed of 0.3 billion km/s, how long would it take for light to travel this distance in sec?
B	How many b are in 3 B?	If a document is 8,000 b in size, how many B does it occupy?	A 1-*min* high-definition video uses a data rate of 8×10^6 B/sec. How many b does the video consume in total?

Table 5: Questions of three difficulty levels (Q_1, Q_2, Q_3) for units of measure.

and 53% respectively in this format.

To investigate the possible source of this phenomenon, we calculate the correlation between NR and post-redefinition anchored responses per LLM. Averaged results for all LLMs are presented in Table 7, revealing an intriguing pattern: for Q_1 and Q_2 levels, the correlation is either weak or negative. A negative correlation indicates that LLMs performing well in NR cases tend to anchor less. This suggests that when reasoning is of easy or medium level, LLMs are less likely to adhere rigidly to their default knowledge. This serves as a sanity check, confirming that LLMs understand the redefinition task and that anchoring rates are not due to prompting deficiencies. However, this trend reverses in the most challenging Q_3 level, particularly in the *swapping* cases. In these instances, a strong positive correlation is evident, implying that LLMs that originally perform well (thus are potent reasoners) tend to disregard the redefinition prompt and respond as they would in the NR case. This striking observation suggests that more capable reasoners anchor more on their prior knowledge. This pattern holds across both FF and MC formats, as well as different prompt types (more results in App. B).

Inverse trends Anchoring is not only related to the per LLM reasoning capabilities, but also to the

Model	R_a3						R_s2					
	Q_1		Q_2		Q_3		Q_1		Q_2		Q_3	
	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC
Mistral7B	**33.33**	**46.67**	**33.33**	**26.67**	26.67	40.0	33.33	53.33	13.33	33.33	26.67	20.0
Mixtral8x7B	33.33	33.33	26.67	26.67	20.0	33.33	26.67	46.67	40.0	**53.33**	46.67	**73.33**
Mistral Large (123B)	33.33	20.0	26.67	26.67	**53.33**	**66.67**	**66.67**	**53.33**	**46.67**	40.0	**73.33**	66.67
Llama8B	0.0	**26.67**	0.0	**26.67**	13.33	33.33	20.0	13.33	**26.67**	40.0	20.0	20.0
Llama70B	**6.67**	13.33	0.0	0.0	13.33	40.0	33.33	46.67	13.33	**46.67**	33.33	73.33
Llama405B	0.0	0.0	0.0	13.33	**26.67**	**53.33**	**26.67**	**46.67**	6.67	20.0	**53.33**	**93.33**
Titan lite	13.33	20.0	20.0	20.0	0.0	40.0	40.0	33.33	20.0	33.33	6.67	26.67
Titan express	20.0	**26.67**	13.33	13.33	**20.0**	13.33	40.0	**53.33**	**20.0**	20.0	33.33	**26.67**
Titan large	**26.67**	20.0	**20.0**	6.67	13.33	**40.0**	**60.0**	40.0	13.33	**33.33**	**33.33**	20.0
Command r	0.0	6.67	**20.0**	**33.33**	**26.67**	**53.33**	**53.33**	13.33	20.0	6.67	**33.33**	**46.67**
Command r +	6.67	13.33	0.0	13.33	13.33	26.67	13.33	20.0	26.67	6.67	33.33	26.67
Command light text	6.67	13.33	13.33	20.0	0.0	40.0	13.33	20.0	**26.67**	**20.0**	13.33	13.33
Command text	**13.33**	**20.0**	6.67	6.67	6.67	26.67	40.0	**26.67**	13.33	26.67	13.33	33.33
Claude opus	13.33	0.0	6.67	6.67	33.33	**46.67**	**46.67**	**40.0**	20.0	**26.67**	53.33	73.33
Claude instant	0.0	13.33	13.33	**20.0**	26.67	46.67	33.33	20.0	33.33	40.0	46.67	60.0
Claude haiku	20.0	13.33	6.67	0.0	20.0	20.0	26.67	6.67	20.0	20.0	40.0	53.33
Claude sonnet v1	26.67	13.33	0.0	13.33	13.33	33.33	33.33	40.0	20.0	20.0	**60.0**	**73.33**
Claude sonnet v2	**26.67**	**13.33**	**20.0**	0.0	**46.67**	40.0	13.33	40.0	**33.33**	20.0	40.0	66.67

Table 6: Anchoring response rate for all LLMs tested using ZS prompting for the most difficult in *assignment* (R_a3) and *swapping* (R_s2) redefinitions. The highest anchoring rate for each LLM family is marked in **bold**.

Level	R_a1	R_a2	R_a3	R_s1	R_s2
Free-Form (FF)					
Q_1	-0.458	-0.071	0.008	0.199	-0.016
Q_2	-0.502	-0.573	-0.472	0.107	0.019
Q_3	0.489	0.237	0.292	0.666	0.668
Multiple Choice (MC)					
Q_1	-0.642	-0.4	-0.344	-0.052	0.025
Q_2	-0.275	-0.316	-0.245	0.41	0.151
Q_3	-0.063	0.457	0.081	0.666	0.75

Table 7: Correlation between average NR correct response rate with anchored response rate for each redefinition and question level in ZS setup. Cells in pink indicate a **high positive correlation** (> 0.3), while cells in green indicate a **high negative correlation** (< -0.3).

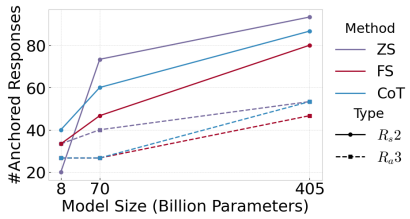


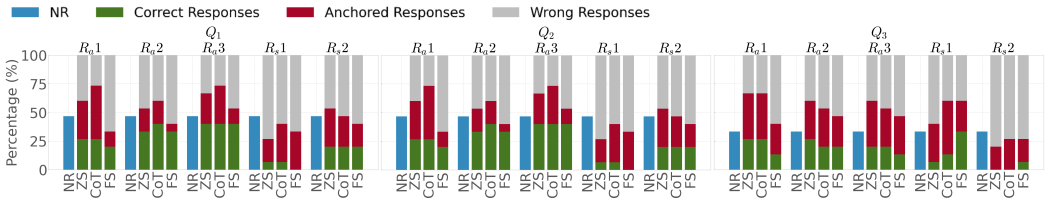
Figure 2: Number of anchored responses for models of varying sizes in the Llama family (MC response format).

parameter size of the LLM itself. Even though larger models achieve higher correct response rate across redefinition levels, staying on par with their reasoning capabilities in the NR case, the anchoring rate also rises as LLM size increases. This indicates that larger models struggle to redefine well-known concepts, and instead rely on their existing knowledge. This is evident from Table 8, which shows the number of correct responses in the NR case versus the anchoring rate. For example, in the case of Llama, the 405B model anchors significantly when solving Q_3 questions post-redefinition compared to the smaller Llama70B, suggesting that the larger variant is less capable as a reasoner. The same trend holds for Mixtral8x7B and Mistral Large (123B): for the latter, anchoring is even higher compared to correct NR responses in the Q_3 level, meaning that the LLM provides the originally correct answer in the redefined problem (when this response is *incorrect*) more frequently than in the NR case (when the answer is *correct*). This unexpected behavior of Llama is further investigated under the MC response format for different prompting methods, focusing on the hardest redefinition (R_a3, R_s2) and question levels (Q_3). As illustrated in Figure 2, the anchoring rate rises with model scale, verifying the *inverse scaling trend*. The same holds for Mistral, as shown in Figure 3, which illustrates the performance of Mistral 7B and Large. Once again, the larger model consistently exhibits a much higher anchoring rate, regardless of the redefinition type or the difficulty of the question—sometimes even exceeding its performance in the NR case.

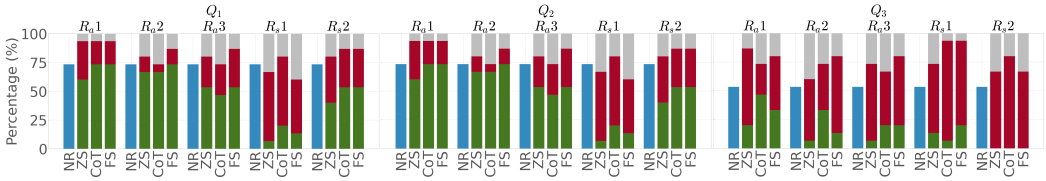
Response format Figure 4 presents results from two LLM families of known parameter count, Mistral and Llama, with varying sizes, for all redefinition levels on Q_3 questions, for FF and MC formats. The MC format is associated with higher anchoring rates (e.g., 73.33% and 93.33% for Llama 70B and 405B respectively) compared to FF responses (33.33% and 53.33%). This is rather expected, since the LLM is exposed to the default value of a constant, creating a conflict between memorization

Model	R_a3						R_s2					
	Q_1		Q_2		Q_3		Q_1		Q_2		Q_3	
	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF
Mistral7B	66.67	33.33	46.67	33.33	33.33	26.67	66.67	33.33	46.67	13.33	33.33	26.67
Mixtral8x7B	100.0	33.33	66.67	26.67	66.67	20.0	100.0	26.67	66.67	40.0	66.67	46.67
Mistral Large (123B)	93.33	33.33	73.33	26.67	53.33	53.33	93.33	66.67	73.33	46.67	53.33	73.33
Llama8B	80.0	0.0	80.0	0.0	53.33	13.33	80.0	20.0	80.0	26.67	53.33	20.0
Llama70B	93.33	6.67	80.0	0.0	80.0	13.33	93.33	33.33	80.0	13.33	80.0	33.33
Llama405B	93.33	0.0	86.67	0.0	73.33	26.67	93.33	26.67	86.67	6.67	73.33	53.33

Table 8: Correct response rate without redefinition (NR) versus post-redefinition anchoring rate in the free-form (FF) format, for LLMs with known sizes using ZS prompting. Colored cells indicate elevated anchoring with LLM scale.

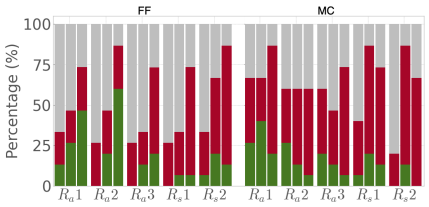


(a) Response breakdown for Mistral 7B before and after constant redefinitions.

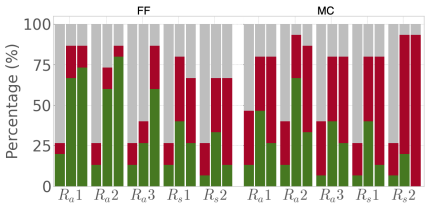


(b) Response breakdown for Mistral Large (123B) before and after constant redefinitions.

Figure 3: Comparison of Mistral 7B and Mistral Large responses on the MC response format.



(a) Response breakdown for Mistral models.



(b) Response breakdown for Llama models.

Figure 4: Results for the different Mistral and Llama models on Q_3 questions using ZS prompting. The order of the bars per redefinition type/level corresponds to increasing model size. The color coding is the same as in Figure 3.

and instruction. The high probability the default value holds triggers the LLM to anchor to it, something that is not applicable in the FF case.

Assignment vs swapping There is a clear distinction between R_a (*assignment*) and R_s (*swapping*) cases: Swapping causes the LLMs to respond with the original constant value more frequently; we hypothesize that this occurs because the LLM's memory is triggered, associating both constants with their default values and thus more readily ignoring redefinition. Notably, this behavior remains consistent across all prompting methods tested.

The influence of prompting Figure 5 highlights the role of prompting in driving the anchoring rate of different LLMs and prompting techniques on Q_3 questions and the hardest R_s2 redefinition level regarding *swapping*. Interestingly, CoT prompting does not help LLMs (even larger ones) avoid anchoring or force them to follow the redefinition task. Instead, FS prompting proves to be more effective in most cases, with 50% of the LLMs tested achieving the minimum anchoring rate using FS. Certain LLMs, such as Mixtral 8x7B/Large, Titan Large, and Claude Haiku, exhibit a significant variance between the maximum and minimum number of anchored responses depending on the prompting technique used. However, this is not a consistent pattern, as most LLMs have a comparable anchoring rate across different techniques. Specifically, the average difference between the maximum and minimum anchoring rate for all LLMs is 16.29 ± 9.22%, indicating that prompting generally has a relatively *small impact* on this phenomenon. Simi-

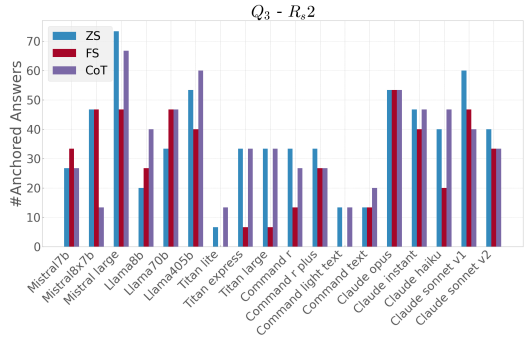


Figure 5: Comparison of the anchored response rate for Q_3 questions in the R_{s2} redefinition level for all LLMs.

lar behaviors are observed for the other redefinition and question levels.

Refusal to respond In some cases, a portion of the LLMs' completely wrong responses does not stem from reasoning inability, but rather from their refusal to perform the redefinition. By refusing to respond, the LLM showcases robustness, since it cannot be misled by a possibly malicious redefinition; however, this behavior obscures the LLM's actual reasoning abilities. To quantify this, we measure LLM refusal rates by categorizing wrong responses into two groups: (i) **actually wrong** and (ii) **redefinition refusal** responses. Table 9 presents the average refusal rate among wrong responses for all question levels. It is evident that *LLM refusal rates vary significantly*, with LLama and Mistral emerging as the LLM families with higher refusal rates. Also, larger models refuse less often, indicating a *false confidence towards providing a response*. Ultimately, refusal is primarily related to LLM scale rather than NR reasoning abilities, with correlations between refusal rate and NR accuracy being weak (0.144 and 0.039 on average for the FF and MC response formats respectively).

Furthermore, prompting techniques play a crucial role in refusal rates, with FS mitigating refusal the most. This result is intuitive, as the LLM is exposed to more examples containing redefinitions in its input, making it less likely to refuse the task. Additional results are provided in App. D.

4.2 Results on units of measure redefinition

The findings on the redefinition of units of measure align with those of constants, also revealing a strong inverse trend: larger models (e.g., Mistral Large, Llama405B) consistently exhibit a higher anchoring rate compared to their smaller counterparts (Mistral7B, Llama8B), regardless of the response format, redefinition level, or prompting

Model	Prompt	FF	MC
Mistral7B	ZS	6.57 ± 11.99	13.34 ± 18.07
	CoT	5.63 ± 8.89	15.62 ± 16.45
	FS	**3.7 ± 7.58**	**10.07 ± 15.25**
Mixtral8x7B	ZS	18.0 ± 22.8	8.61 ± 16.97
	CoT	9.22 ± 16.82	15.5 ± 17.63
	FS	**10.98 ± 17.03**	**5.95 ± 18.79**
Mistral Large	ZS	16.33 ± 33.69	1.67 ± 6.24
	CoT	**8.33 ± 18.51**	**0 ± 0**
	FS	14.35 ± 26.96	1.33 ± 4.99
Llama8B	ZS	55.54 ± 24.37	40.05 ± 18.58
	CoT	35.25 ± 23.33	32.89 ± 23.21
	FS	**2.41 ± 6.64**	**0 ± 0**
Llama70B	ZS	38.66 ± 29.92	5.56 ± 14.49
	CoT	9.17 ± 17.36	13.33 ± 27.35
	FS	**0 ± 0**	**0 ± 0**
Llama405B	ZS	1.33 ± 4.99	**0 ± 0**
	CoT	**0 ± 0**	**0 ± 0**
	FS	**0 ± 0**	**0 ± 0**
Titan lite	ZS	**1.56 ± 3.19**	**0 ± 0**
	CoT	3.03 ± 5.66	**0 ± 0**
	FS	2.54 ± 5.39	**0 ± 0**
Titan express	ZS	0.56 ± 2.08	**0 ± 0**
	CoT	1.9 ± 7.13	**0 ± 0**
	FS	**0 ± 0**	**0 ± 0**
Titan large	ZS	2.0 ± 5.42	**0 ± 0**
	CoT	**0 ± 0**	**0 ± 0**
	FS	**0 ± 0**	**0 ± 0**
Command text	ZS	3.33 ± 9.03	**0 ± 0**
	CoT	**0 ± 0**	**0 ± 0**
	FS	0.83 ± 3.12	**0 ± 0**
Claude instant	ZS	1.69 ± 4.36	**0 ± 0**
	CoT	**0 ± 0**	**0 ± 0**
	FS	4.07 ± 12.58	**0 ± 0**
Claude sonnet v2	ZS	20.48 ± 26.25	4.83 ± 9.29
	CoT	14.31 ± 24.39	10.0 ± 27.08
	FS	**8.91 ± 24.75**	**3.17 ± 8.81**

Table 9: Average refusal rates over all question levels (lowest values in **bold** and highest values underlined). We exclude LLMs with zero refusal rate overall.

method. This trend is particularly notable in Q_3 questions. However, anchoring is relatively lower than in the constants case, likely because the actual values of units are not commonly used in calculations, making the LLMs less prone to anchoring. An thorough analysis is presented in App. E.

5 Conclusion

In this work, we thoroughly investigate the redefinition task by prompting LLMs to reason with redefined values of physical constants and units of measure. We uncover several critical patterns in LLMs, showcasing pitfalls of scale, such as decreased reasoning capacity and increased confidence in erroneous answers instead of abstaining. Moreover, we offer extensive insights into how redefinition difficulty, prompting strategies, and response format influence LLMs' propensity to anchor on their prior knowledge rather than reason flexibly.

Limitations

Our redefinitions focus on concept sets with a well-defined number of elements, such as mathematical constants and unit measures, restricting our investigation to closed-world reasoning rather than broader, more generalizable tasks. We opt for such restricted settings in order to evaluate more clearly the impact of redefinition difficulty in conjunction to question difficulty, targeting certain LLM reasoning capabilities. Evaluating additional types of reasoning over redefinitions should consider more broad concept sets to be redefined.

Furthermore, we do not compare LLM performance on redefinition tasks with human performance, following the assertion of McKenzie et al. (2024) that such tasks are generally easy for humans, albeit requiring some effort in complex cases. Conducting a human experiment for direct comparison would be highly challenging due to significant variability in individual knowledge and expertise. Prior exposure to related tasks could further bias results—for example, a physics teacher may solve redefinition problems with ease, whereas others may struggle. Moreover, concentration, memory, motivation, engagement, psychological and environmental factors play a decisive role in human performance, making controlled experimentation significantly difficult and possibly unreliable.

Ethical considerations

The ability to redefine concepts in LLMs presents ethical challenges, particularly in the generation of misleading or deceptive responses. Our study highlights an inherent trade-off between reasoning transparency and model robustness. More robust models resist redefinition by refusing the task, making them less susceptible to manipulation but also limiting their ability to engage in flexible reasoning. Conversely, models that effectively reason with redefined values exhibit greater transparency and adaptability but are more vulnerable to malicious prompts. This duality raises a significant ethical question: should LLMs prioritize strict factual adherence at the cost of reasoning flexibility, or should they remain adaptable at the risk of being misled? Addressing this trade-off is a crucial ethical consideration in the responsible design and deployment of LLMs in general.

References

Thomas Ball, Shuo Chen, and Cormac Herley. 2024. Can we count on llms? the fixed-effect fallacy and claims of gpt-4 capabilities. *ArXiv*, abs/2409.07638.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiaao Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2024. Skills-in-context: Unlocking compositionality in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13838–13890, Miami, Florida, USA. Association for Computational Linguistics.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

Gaël Gendron, Bao Trung Nguyen, Alex Yuxuan Peng, Michael Witbrock, and Gillian Dobbie. 2024. Can large language models learn independent causal mechanisms? In *Conference on Empirical Methods in Natural Language Processing*.

Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. 2024. Puzzle solving using reasoning of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11574–11591, Miami, Florida, USA. Association for Computational Linguistics.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? *Preprint*, arXiv:2306.05836.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *ArXiv*, abs/2402.08955.

Chenxi Li, Yuanhe Tian, Zhaxi Zerong, Yan Song, and Fei Xia. 2024. Challenging large language models with new tasks: A study on their adaptability and robustness. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8140–8162, Bangkok, Thailand. Association for Computational Linguistics.

Siyu Lou, Yuntian Chen, Xiaodan Liang, Liang Lin, and Quanshi Zhang. 2024. Quantifying in-context reasoning effects and memorization effects in llms. *Preprint*, arXiv:2405.11880.

Kyle Mahowald, Anna A. Ivanova, Idan Asher Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28:517–540.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. 2024. Inverse scaling: When bigger isn't better. *Preprint*, arXiv:2306.09479.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2024. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *ArXiv*, abs/2406.02061.

Sumedh Rasal and E. J. Hauer. 2024. Navigating complexity: Orchestrated problem solving with multi-agent llms. *Preprint*, arXiv:2402.16713.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, An-743 drea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Stephen Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cèsar Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Christopher Callison-Burch, Christian Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Daniel H Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hanna Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Narain Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Jane W Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jorg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Josh Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras Ochando, Louis-Philippe Morency, Luca Moschella, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, María José Ramírez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis,

Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michal Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mohit Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, T. MukundVarma, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, P. Milkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphael Milliere, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Samuel Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi S. Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yufang Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.*, 2023.

Claire E. Stevenson, Alexandra Pafford, Han L. J. van der Maas, and Melanie Mitchell. 2024. Can large language models generalize analogy solving like people can? *ArXiv*, abs/2411.02348.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2024a. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. *Preprint*, arXiv:2407.14985.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024b. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024. On memorization of large language models in logical reasoning. *Preprint*, arXiv:2410.23123.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Q_1 exceptions

Regarding the constants redefinition task, we consider some exceptions to the default Q_1: *What is the first -non-zero- digit of {constant}?*. These exceptions are listed in Table 10. Those questions directly trigger the existing knowledge of LLMs, mostly requesting retrieving rather than reasoning on the queried information. For example, it is well known that $i^2 = -1$ and there is no reasoning on the question Q_1*: What is the value of* i^2*?*.

	Q_1
c	How far does light travel in one second?
i	What is the value of i^2?
∞	What is the value of infinity?
zero	What is the absolute value of zero?

Table 10: Exceptions to the typical Q_1 format

B Model Knowledgeability

Tables 11 and 12 show the correlation between performance before redefinition (NR case) and the anchored response rate for each constant redefinition and question level in the FS and CoT prompting setups respectively. The pattern remains the same: in Q_1 and Q_2 question levels, there is little to no correlation or a negative correlation between the two values, whereas in Q_3—particularly in the *swapping* case—the correlation is highly positive. This suggests that in those cases, more knowledgeable models adapt less to the redefinition.

Table 13 presents the number of correct answers in the NR case alongside the percentage of anchored responses for constant redefinitions across all LLMs used in the study.

Level	R_a1	R_a2	R_a3	R_s1	R_s2
Free-Form (FF)					
Q_1	-0.055	-0.129	-0.472	0.235	-0.008
Q_2	-0.283	-0.359	-0.444	0.085	-0.148
Q_3	0.356	0.374	0.492	0.596	0.823
Multiple Choice (MC)					
Q_1	-0.71	-0.624	-0.711	-0.304	-0.28
Q_2	-0.258	-0.473	-0.312	0.441	-0.15
Q_3	0.269	0.589	0.288	0.624	0.694

Table 11: Correlation between model performance before redefinition with the percentage of anchored answers for each type of constant redefinition and question level in FS setup. Cells highlighted in pink indicate a **high positive correlation** (> 0.3), while cells in green indicate a **high negative correlation** (< -0.3).

Level	R_a1	R_a2	R_a3	R_s1	R_s2
Free-Form (FF)					
Q_1	-0.539	-0.542	-0.552	-0.244	-0.319
Q_2	-0.521	-0.626	-0.58	0.143	-0.125
Q_3	0.41	0.116	-0.085	0.71	0.588
Multiple Choice (MC)					
Q_1	-0.529	-0.483	-0.358	-0.17	0.16
Q_2	-0.183	-0.224	-0.202	0.329	-0.044
Q_3	0.134	0.366	0.009	0.679	0.657

Table 12: Correlation between model performance before redefinition with the percentage of anchored answers for each type of constant redefinition and question level in CoT setup. Cells highlighted in pink indicate a **high positive correlation** (> 0.3), while cells in green indicate a **high negative correlation** (< -0.3).

C Inverse Trends

Figure 7 shows the rate of correct answers in the NR task, as well as the rates of correct, wrong, and anchored responses for LLaMA-8B and LLaMA-405B after the constant redefinition task in the MC response format. From this figure, it is observable that the same pattern as in Mistral (Figure 3) emerges, where the number of anchored responses is significantly higher in the larger model. Additionally, Figure 6 shows the number of anchored responses after constants redefinitions for models of varying sizes in the Mistral family in the MC response format. The trend is the same as in Llama models illustrated in the main paper, where larger variants tend to anchor more to their prior knowledge in comparison to the smaller ones, for both R_s1, R_s2 redefinition levels and all prompting techniques. Mixtral7x8B in the ZS prompting setup produces only slightly more anchored responses compared to the larger variant (Mistral Large (123B parameters)). However, this is likely due to the increased performance of the LLM in the NR case, as shown in Table 8. The difference is relatively small, meaning that while there is a measurable increase in anchoring for Mixtral7x8B, it is not substantial enough to indicate a significant deviation from the expected trend.

D Refusal Rates

An important observation stated in the main paper is that completely wrong responses include instances where the LLM actively refuses to respond to the redefined problem, as analyzed before.

An analysis of the most interesting cases regarding wrong responses is presented in Figures 8, 9, 10, 11 for selected LLMs. For example, in the case of Llama8B (Figure 8) the refusal rate diminishes

Model	R_a3						R_s2					
	Q_1		Q_2		Q_3		Q_1		Q_2		Q_3	
	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF
Mistral7B	66.67	33.33	46.67	33.33	33.33	26.67	66.67	33.33	46.67	13.33	33.33	26.67
Mixtral8x7B	100.0	33.33	66.67	26.67	66.67	20.0	100.0	26.67	66.67	40.0	66.67	46.67
Mistral Large (123B)	93.33	33.33	73.33	26.67	53.33	53.33	93.33	66.67	73.33	46.67	53.33	73.33
Llama8B	80.0	0.0	80.0	0.0	53.33	13.33	80.0	20.0	80.0	26.67	53.33	20.0
Llama70B	93.33	6.67	80.0	0.0	80.0	13.33	93.33	33.33	80.0	13.33	80.0	33.33
Llama405B	93.33	0.0	86.67	0.0	73.33	26.67	93.33	26.67	86.67	6.67	73.33	53.33
Titan lite	46.67	13.33	20.0	20.0	20.0	0.0	46.67	40.0	20.0	20.0	20.0	6.67
Titan express	73.33	20.0	33.33	13.33	26.67	20.0	73.33	40.0	33.33	20.0	26.67	33.33
Titan large	66.67	26.67	33.33	20.0	26.67	13.33	66.67	60.0	33.33	13.33	26.67	33.33
Command r	86.67	0.0	33.33	20.0	40.0	26.67	86.67	53.33	33.33	20.0	40.0	33.33
Command r +	93.33	6.67	66.67	0.0	66.67	13.33	93.33	13.33	66.67	26.67	66.67	33.33
Command light text	60.0	6.67	6.67	13.33	0.0	0.0	60.0	13.33	6.67	26.67	0.0	13.33
Command text	53.33	13.33	40.0	6.67	26.67	6.67	53.33	40.0	40.0	13.33	26.67	13.33
Claude opus	100.0	13.33	80.0	6.67	80.0	33.33	100.0	46.67	80.0	20.0	80.0	53.33
Claude instant	86.67	0.0	33.33	13.33	46.67	26.67	86.67	33.33	33.33	33.33	46.67	46.67
Claude haiku	100.0	20.0	73.33	6.67	66.67	20.0	100.0	26.67	73.33	20.0	66.67	40.0
Claude sonnet v1	100.0	26.67	93.33	0.0	86.67	13.33	100.0	33.33	93.33	20.0	86.67	60.0
Claude sonnet v2	73.33	26.67	60.0	20.0	40.0	46.67	73.33	13.33	60.0	33.33	40.0	40.0

Table 13: The percentage of correct responses with no redefinition (NR) and the anchored responses after constant redefinitions regarding free-form (FF) responses and using ZS prompting.

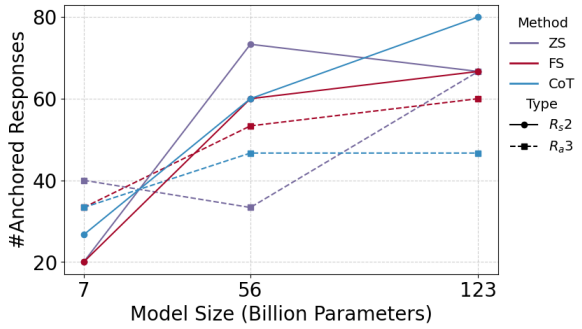


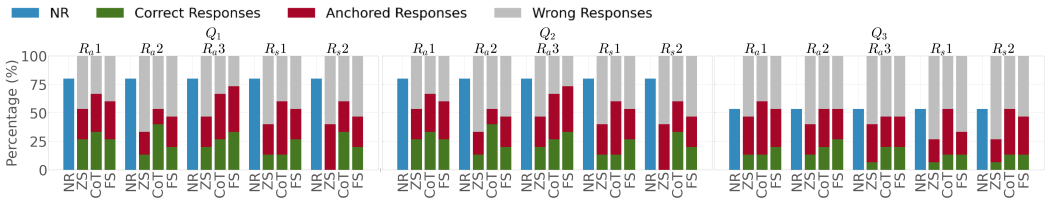
Figure 6: Anchored response rate after constants redefinitions for LLMs of varying sizes in the Mistral family under the MC response format, harnessing different prompting techniques on the hardest redefinition levels.

as questions get harder. This reveals an overconfidence in responding instead of abstaining, leading to a problematic behavior, since Llama8B achieves very few correct responses in all question levels, and especially in harder ones (Q_3), as indicated in Figure 7a. the difference between FF and MC response formats lies in the elevated number of blank responses in the MC case. This is because Llama8B suffers when prompted to select one of the predefined options, indicating high uncertainty. Nevertheless, when prompted to respond without restrictions, empty responses are almost non-existent, revealing another sense of overconfidence tied to response format.

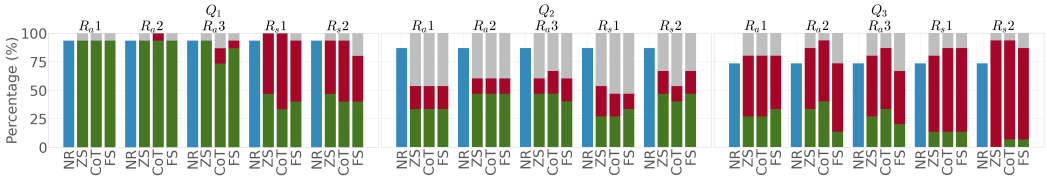
Conversely, Llama70B (Figure 9) refrains from empty responses, especially in harder cases, revealing its lower uncertainty in generating a response. Contrary to its smaller counterpart, its refusal rates decrease as questions get harder, leading to an increased number of wrong responses (reaching up to 100% in some cases) over refusals.

A mixed behavior is presented in the case of Mixtral8x7B (Figure 10), where refusal decreases in the FF response format, while it increases in the MC format. This behavior denotes that when Mixtral8x7B is exposed to a limited set of options, it elevates its resistance in redefining constants, possibly detecting the presented conflict between memorization and instruction. On the other hand, when generation is unrestricted, as in the FF case, its denial becomes significantly diminished, resulting in erroneous responses more often than not. Ultimately, in this case, response format is of outmost importance in defining the trade-off between response refusal and erroneous generations.

Finally, mixed patterns also occur in the case of Claude Sonnet v2 (Figure 11), where alternating patterns between 100% refusal or 100% wrong responses are revealed (as in the Q_1 question level). Therefore, this model is rather unpredictable in whether it prefers to deny the task or respond erroneously, since there is no obvious reason behind this diverging behavior. Contrary to the aforementioned Mixtral8x7 case, Claude Sonnet v2 presents more wrong responses in the MC case in comparison to FF responses. That means that for some unexpected reason, it refuses to answer within an unrestricted setting, but results in wrong responses when presented with a limited number of options.



(a) Response breakdown for Llama8B before and after constant redefinitions.



(b) Response breakdown for Llama405B before and after constant redefinitions.

Figure 7: Comparison of Llama8B and Llama405B responses on the MC response format.

This behavior reveals that Claude Sonnet v2 confidently performs redefinitions with ease and without much resistance, but fails to solve the redefined problem overall.

Other than that, Tables 14, 15, and 16 present the proportion of incorrect responses attributed to the LLM's refusal to respond to the constants redefinition task over all completely wrong responses for all LLMs together. These Tables report refusal rates for each LLM, prompting method, and redefinition level regarding redefinitions across all Q_1, Q_2, and Q_3 question levels respectively.

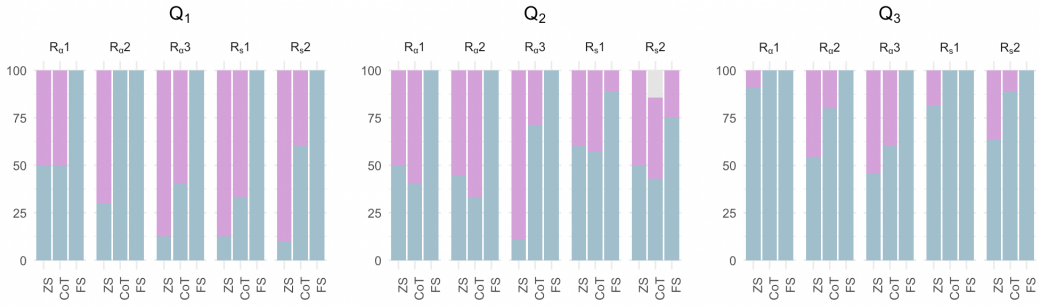
The results indicate that models such as Command r+, Command light, and Claude Haiku consistently tend to provide responses, ignoring their validity, therefore exhibiting *no refusal instances*. This overconfidence is problematic in practice, even though useful in our experimentation, since reasoning shortcomings are exposed. In contrast, models such as Llama, Mistral, and Claude sonnet v2 occasionally decline to generate a response when faced with the redefined task, possibly acknowledging their intrinsic inability to answer properly. This observation showcases that those LLMs that prefer to abstain from responding are more robust, since a redefinition prompt could act as a malicious adversarial attack that aims to mislead the LLM towards generating invalid responses. On the other hand, this deliberate action prevents them to reveal their reasoning capabilities, once again verifying the trade-off between robustness and evident reasoning.

Additionally, the type of prompting significantly influences the LLM's refusal rate. Specifically, models exhibit lower refusal rates in the FS setup compared to the ZS and CoT configurations. We assume that this is because LLMs may 'feel' overconfident when exposed to FS exemplars that clearly showcase the redefinition task to be performed, overriding their inherent inability to actually and properly reason over redefined concepts.

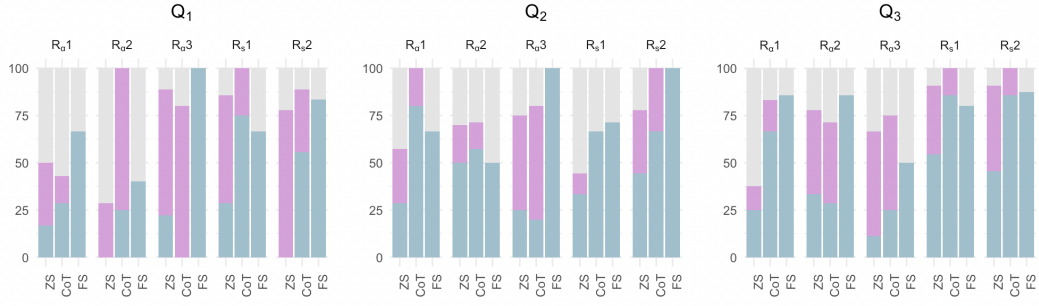
To further investigate LLM anchoring more accurately, we calculate the rate of anchored responses *only* in cases where the LLM indeed attempts to solve the problem, excluding refusal cases. Table 17 presents this pure refusal rate for models in the ZS prompting setup, focusing on the most challenging redefinitions in *assignment* (R_a3) and *swapping* (R_s2) cases. We exclude LLMs where no refusals occurred, as their results are identical to those reported in Table 6.

E Results on units of measure redefinition

An overview of response accuracy is presented in Table 18, where we consider the hardest redefinitions corresponding to R_a3 and R_s2 for *assignment* and *swapping* respectively, as well as all three question levels, together with FF and MC response formats regarding units of measurement redefinitions. Additionally, Table 19 presents the number of correct responses in the NR case alongside the anchoring rate for FF responses regarding units redefinitions. Anchoring is less prominent for some LLMs in comparison to their anchored responses in the constants redefinition task; for instance, Command r+, light text, text achieve even 0% anchoring in some cases, even in Q_3 questions over the hardest R_a3 unit redefinitions. Nevertheless, anchoring still persists in many instances, with large rates concerning models in the Mistral family for the hardest question and redefinition levels. Moreover, Titan models present high anchoring even for R_a2



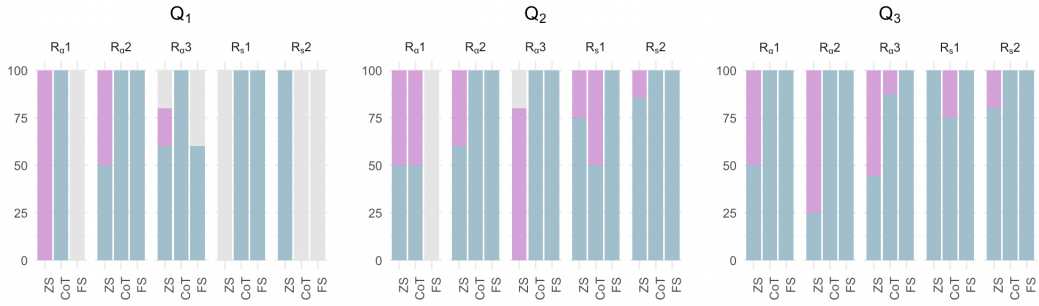
(a) Response breakdown for Llama8B FF responses.



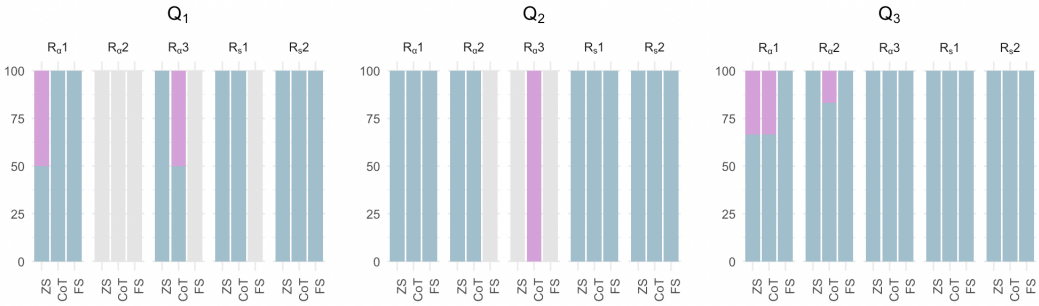
(b) Response breakdown for Llama8B MC responses.

Figure 8: Completely wrong responses breakdown for Llama8B. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses

.



(a) Response breakdown for Llama70B FF responses.

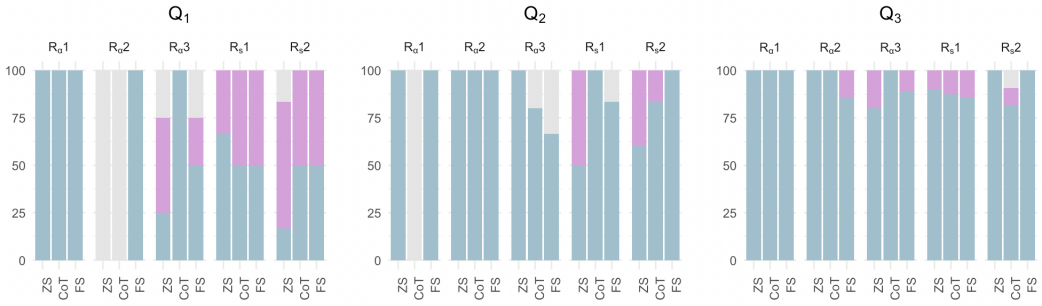


(b) Response breakdown for Llama70B MC responses.

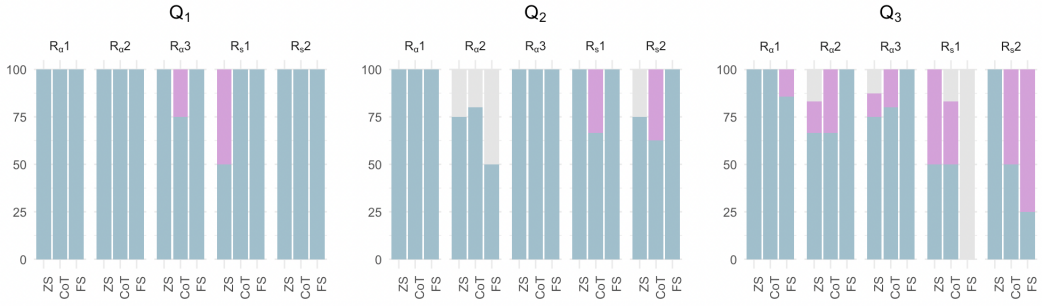
Figure 9: Completely wrong responses breakdown for Llama70B. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses

.

unit redefinitions, even in the easier Q_1 level. Surprisingly, anchoring for Titan models reduces as questions and redefinitions become harder, but this does not indicate an improvement in producing correct responses and therefore an advancement in reasoning capability; instead, the anchoring reduction



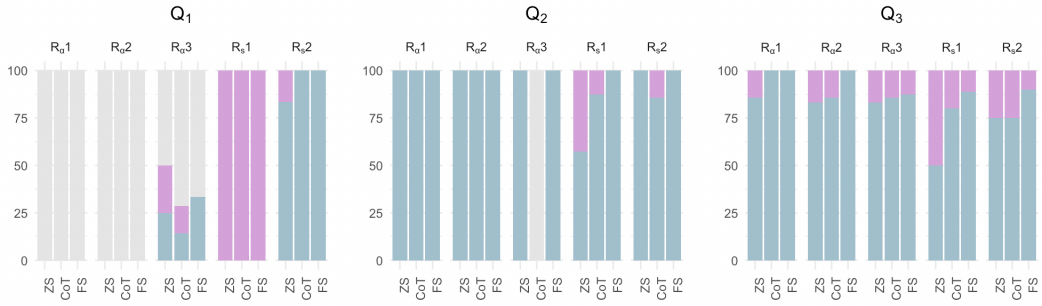
(a) Response breakdown for Mixtral8x7 FF responses.



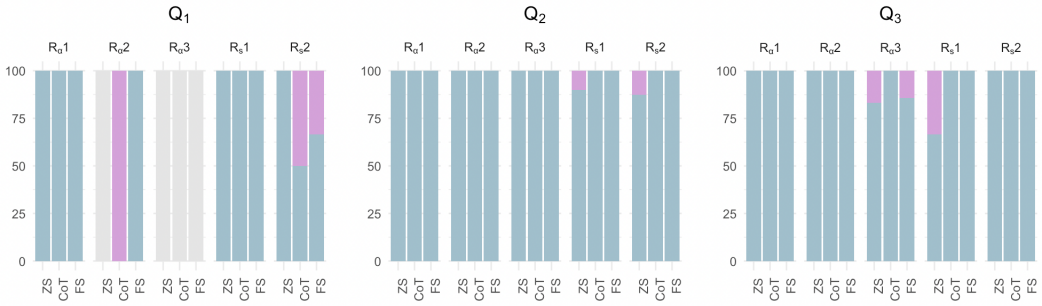
(b) Response breakdown for Mixtral8x7 MC responses.

Figure 10: Completely wrong responses breakdown for Mixtral8x7. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses

.



(a) Response breakdown for Claude Sonnet v2 FF responses.



(b) Response breakdown for Claude Sonnet v2 MC responses.

Figure 11: Completely wrong responses breakdown for Claude Sonnet v2. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses

.

is attributed to the generation of more completely wrong responses, indicating those models' inability of solving the unit of measure redefinition task appropriately.

Figure 12 shows the results of the different Mistral and Llama models for the Q_3 question level

Model	Prompt	FF					MC				
		R_s1	R_s2	R_s3	R_a1	R_a2	R_s1	R_s2	R_s3	R_a1	R_a2
Mistral7B	ZS	0.0	0.0	0.0	16.67	44.44	40.0	0.0	50.0	50.0	0.0
	CoT	0.0	0.0	0.0	25.0	11.11	0.0	40.0	0.0	50.0	0.0
	FS	14.29	0.0	18.18	0.0	0.0	0.0	16.67	50.0	0.0	0.0
Mixtral8x7B	ZS	0.0	0.0	50.0	33.33	66.67	0.0	0.0	0.0	50.0	0.0
	CoT	0.0	0.0	0.0	50.0	50.0	0.0	0.0	25.0	0.0	0.0
	FS	0.0	0.0	25.0	50.0	50.0	0.0	0.0	0.0	0.0	0.0
Mistral Large	ZS	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	66.67	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	100.0	50.0	0.0	0.0	0.0	0.0	0.0
Llama8B	ZS	50.0	70.0	87.5	87.5	90.0	33.33	28.57	66.67	57.14	77.78
	CoT	50.0	0.0	60.0	66.67	40.0	14.29	75.0	80.0	25.0	33.33
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama70B	ZS	100.0	50.0	20.0	0.0	0.0	50.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	50.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama405B	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Titan lite	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	14.29	0.0	0.0	16.67	0.0	0.0	0.0	0.0	0.0
Titan express	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	28.57	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Titan large	ZS	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command r	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command r plus	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command light text	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command text	ZS	0.0	0.0	0.0	0.0	0.0	0.0	16.67	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	12.5	0.0	0.0	0.0	0.0	0.0
Claude opus	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude instant	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	50.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude haiku	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude sonnet v1	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude sonnet v2	ZS	0.0	0.0	25.0	100.0	16.67	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	14.29	100.0	0.0	0.0	100.0	0.0	0.0	50.0
	FS	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	33.33

Table 14: The refusal rate for each LLM, method, and type of constants redefinitions for Q_1 questions.

in the ZS prompting setup for units redefinitions. The conclusions are similar to those for the redefinition of constants, where the number of anchored responses is significantly higher in the MC setup compared to the FF setup –a rather expected pattern, since LLMs are exposed to the default response. Additionally, once again, it is observable that the larger models are more prone to providing anchored responses compared to the smaller ones, regardless the response format and the model family.

Furthermore, Figures 13 and 14 illustrate the results of Mistral7B and Mistral Large, as well as Llama8B and Llama 405B, respectively, regarding

Model	Prompt	FF					MC				
		R_s1	R_s2	R_s3	R_a1	R_a2	R_s1	R_s2	R_s3	R_a1	R_a2
Mistral7B	ZS	0.0	0.0	0.0	20.0	8.33	0.0	0.0	20.0	9.09	14.29
	CoT	0.0	0.0	0.0	14.29	25.0	0.0	0.0	25.0	0.0	25.0
	FS	0.0	0.0	0.0	0.0	0.0	10.0	33.33	0.0	30.0	11.11
Mixtral8x7B	ZS	0.0	0.0	0.0	50.0	40.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	16.67	0.0	0.0	0.0	33.33	37.5
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral Large	ZS	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	28.57	0.0	0.0	0.0	0.0	0.0
Llama8B	ZS	50.0	55.56	88.89	40.0	50.0	28.57	20.0	50.0	11.11	33.33
	CoT	60.0	66.67	28.57	42.86	42.86	20.0	14.29	60.0	0.0	33.33
	FS	0.0	0.0	0.0	11.11	25.0	0.0	0.0	0.0	0.0	0.0
Llama70B	ZS	50.0	40.0	80.0	25.0	14.29	0.0	0.0	0.0	0.0	0.0
	CoT	50.0	0.0	0.0	50.0	0.0	0.0	0.0	100.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama405B	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Titan lite	ZS	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	7.14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Titan express	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Titan large	ZS	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command r	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command r+	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command light text	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command text	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude opus	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude instant	ZS	0.0	0.0	0.0	11.11	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	11.11	0.0	0.0	0.0	0.0	0.0	0.0
Claude haiku	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude sonnet v1	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude sonnet v2	ZS	0.0	0.0	0.0	42.86	0.0	0.0	0.0	0.0	10.0	12.5
	CoT	0.0	0.0	0.0	12.5	14.29	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 15: The refusal rate for each LLM, method, and type of constants redefinition for Q_2 questions.

units of measure redefinitions. Once again, Mistral7B tends to provide *fewer anchored responses* compared to its larger counterpart. The same trend holds for Llama, although for Q_1 and Q_2 responses, the increase in anchoring is less pronounced for the larger model.

Additionally, the performance of the larger LLama405B model in the NR case is excellent in the Q_1 question level, achieving a response accuracy close to 100% in most cases, denoting that this model is adequately knowledgeable regarding the default meanings of units of measure. For Q_2 questions in Llama405B, an interesting pattern emerges. The number of correct responses in the NR task

Model	Prompt	FF					MC				
		R_s1	R_s2	R_s3	R_a1	R_a2	R_s1	R_s2	R_s3	R_a1	R_a2
Mistral7B	ZS	0.0	0.0	9.09	0.0	0.0	0.0	0.0	0.0	0.0	16.67
	CoT	0.0	0.0	0.0	0.0	9.09	40.0	14.29	14.29	16.67	9.09
	FS	0.0	0.0	23.08	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mixtral8x7B	ZS	0.0	0.0	20.0	10.0	0.0	0.0	16.67	12.5	50.0	0.0
	CoT	0.0	0.0	0.0	12.5	9.09	0.0	33.33	20.0	33.33	50.0
	FS	0.0	14.29	11.11	14.29	0.0	14.29	0.0	0.0	0.0	75.0
Mistral Large	ZS	0.0	0.0	25.0	0.0	0.0	0.0	0.0	0.0	25.0	0.0
	CoT	0.0	0.0	25.0	0.0	33.33	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	20.0	0.0	16.67	0.0	0.0	0.0	0.0	20.0
Llama8B	ZS	9.09	45.45	54.55	18.18	36.36	12.5	44.44	55.56	36.36	45.45
	CoT	0.0	20.0	40.0	0.0	11.11	16.67	42.86	50.0	14.29	14.29
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama70B	ZS	50.0	75.0	55.56	0.0	20.0	33.33	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	12.5	25.0	0.0	33.33	16.67	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama405B	ZS	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Titan lite	ZS	0.0	0.0	6.67	6.67	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	7.69	7.69	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Titan express	ZS	0.0	0.0	0.0	8.33	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Titan large	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command r	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command r+	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command light text	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Command text	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.33
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude opus	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude instant	ZS	0.0	0.0	0.0	14.29	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude haiku	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude sonnet v1	ZS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CoT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Claude sonnet v2	ZS	14.29	16.67	16.67	50.0	25.0	0.0	0.0	16.67	33.33	0.0
	CoT	0.0	14.29	14.29	20.0	25.0	0.0	0.0	0.0	0.0	0.0
	FS	0.0	0.0	12.5	11.11	10.0	0.0	0.0	14.29	0.0	0.0

Table 16: The refusal rate for each LLM, method, and type of constants redefinition for Q_3 questions.

remains close to 100%, indicating that the model is also an excellent reasoner in this difficulty level. However, when units are redefined, the model's accuracy *declines considerably*, coinciding with a noticeable increase in the number of anchored responses. Therefore, Llama405B exploits memorized patterns to be able to handle unit redefinitions, even though memorization almost useless in the Q_1 level. The anchoring rate further increases in the Q_3 level, even though the NR correct response rate (and therefore the model's reasoning ability in the default setting) is decreased in comparison to the easier question levels.

Lastly, Tables 20, 21, and 22 present the cor-

Model	R_a3						R_s2					
	Q_1		Q_2		Q_3		Q_1		Q_2		Q_3	
	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC
Mistral7B	33.33	**63.64**	**33.33**	**31.25**	28.57	40.0	47.36	53.33	14.37	36.84	26.67	23.08
Mixtral8x7B	**50.0**	33.33	26.67	26.67	23.81	36.36	52.18	46.67	52.63	**53.33**	46.67	**73.33**
Mistral Large	33.33	20.0	26.67	26.67	**60.37**	**66.67**	**100.0**	**53.33**	**52.24**	40.0	**73.33**	66.67
Llama8B	0.0	**52.18**	0.0	**42.11**	25.28	52.94	**71.43**	40.9	**42.11**	**50.0**	28.2	31.43
Llama70B	**8.2**	13.33	0.0	0.0	25.71	40.0	33.33	46.67	15.21	46.67	38.46	73.33
Llama405B	0.0	0.0	0.0	13.33	**26.67**	**53.33**	26.67	**46.67**	6.67	20.0	**58.82**	**93.33**
Command text	**13.33**	**20.0**	6.67	6.67	6.67	26.67	40.0	**26.67**	13.33	26.67	13.33	42.85
Claude v2	**32.66**	**13.33**	**20.0**	0.0	**51.22**	44.45	15.58	40.0	**33.33**	22.22	47.06	66.67

Table 17: The percentage of anchored responses for the models in the ZS setup for the most difficult constants redefinitions in *assignment* (R_a3) and *swapping* (R_s2). The highest number for each model family is presented in **bold**. We exclude models where no refusals occurred, as their results are identical to those in Table 6.

Model	R_a2						R_a3					
	Q_1		Q_2		Q_3		Q_1		Q_2		Q_3	
	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC
Mistral7B	0.0	37.5	25.0	25.0	18.75	56.25	**62.5**	25.0	**31.25**	**37.5**	31.25	25.0
Mixtral8x7B	**6.25**	31.25	**31.25**	**37.5**	31.25	37.5	6.25	**31.25**	6.25	31.25	**31.25**	**50.0**
Mistral Large	0.0	**37.5**	6.25	37.5	12.5	56.25	0.0	25.0	12.5	**37.5**	12.5	43.75
Llama8B	0.0	**25.0**	**6.25**	**31.25**	12.5	31.25	**6.25**	**31.25**	**12.5**	**50.0**	**25.0**	50.0
Llama70B	0.0	6.25	**6.25**	**31.25**	**25.0**	**56.25**	0.0	18.75	0.0	**50.0**	12.5	**62.5**
Llama405B	0.0	0.0	0.0	**31.25**	12.5	37.5	0.0	0.0	6.25	25.0	**25.0**	31.25
Titan lite	6.25	**25.0**	12.5	**31.25**	12.5	**25.0**	25.0	**31.25**	25.0	12.5	0.0	18.75
Titan express	18.75	**25.0**	**25.0**	18.75	12.5	**25.0**	**43.75**	25.0	31.25	12.5	6.25	18.75
Titan large	**31.25**	12.5	12.5	**31.25**	**18.75**	**25.0**	25.0	12.5	**37.5**	**31.25**	6.25	**25.0**
Command r	**12.5**	18.75	**12.5**	**31.25**	25.0	18.75	6.25	25.0	**12.5**	18.75	**12.5**	31.25
Command r+	6.25	**43.75**	0.0	25.0	**37.5**	**50.0**	**6.25**	**31.25**	0.0	**31.25**	0.0	25.0
Command light text	6.25	12.5	0.0	25.0	6.25	25.0	12.5	25.0	6.25	31.25	0.0	**50.0**
Command text	12.5	12.5	12.5	18.75	0.0	18.75	0.0	31.25	12.5	12.5	0.0	43.75
Claude opus	0.0	0.0	0.0	6.25	12.5	25.0	0.0	0.0	0.0	0.0	0.0	6.25
Claude instant	**6.25**	**25.0**	**12.5**	25.0	0.0	**43.75**	0.0	**43.75**	0.0	**37.5**	6.25	**31.25**
Claude haiku	0.0	18.75	0.0	12.5	6.25	31.25	0.0	6.25	0.0	6.25	**18.75**	**31.25**
Claude sonnet v2	0.0	0.0	0.0	12.5	6.25	6.25	0.0	0.0	0.0	6.25	0.0	0.0
Claude sonnet v2 S	**6.25**	18.75	6.25	**31.25**	**18.75**	31.25	**6.25**	0.0	**6.25**	25.0	6.25	12.5

Table 18: The percentage of anchored responses for all LLMs tested under the ZS prompting setup for the most difficult units of measure redefinitions in *assignment* (R_a3) and *swapping* (R_a2). The highest rate for each model family is presented in **bold**.

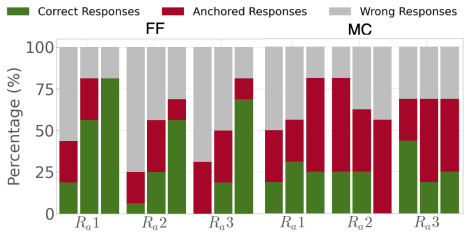
Model	R_a2						R_a3					
	Q_1		Q_2		Q_3		Q_1		Q_2		Q_3	
	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF
Mistral 7B	81.25	0.0	56.25	25.0	43.75	18.75	81.25	62.5	56.25	31.25	43.75	31.25
Mixtral8x7B	87.5	6.25	81.25	31.25	62.5	31.25	87.5	6.25	81.25	6.25	62.5	31.25
Mistral Large	93.75	0.0	93.75	6.25	81.25	12.5	93.75	0.0	93.75	12.5	81.25	12.5
Llama8B	75.0	0.0	56.25	6.25	6.25	12.5	75.0	6.25	56.25	12.5	6.25	25.0
Llama70B	100.0	0.0	81.25	6.25	56.25	25.0	100.0	0.0	81.25	0.0	56.25	12.5
Llama405B	100.0	0.0	93.75	0.0	56.25	12.5	100.0	0.0	93.75	6.25	56.25	25.0
Titan lite	37.5	6.25	18.75	12.5	6.25	12.5	37.5	25.0	18.75	25.0	6.25	0.0
Titan express	75.0	18.75	37.5	25.0	6.25	12.5	75.0	43.75	37.5	31.25	6.25	6.25
Titan large	68.75	31.25	68.75	12.5	25.0	18.75	68.75	25.0	68.75	37.5	25.0	6.25
Command r	75.0	12.5	56.25	12.5	18.75	25.0	75.0	6.25	56.25	12.5	18.75	12.5
Command r+	87.5	6.25	93.75	0.0	81.25	37.5	87.5	6.25	93.75	0.0	81.25	0.0
Command light text	31.25	6.25	6.25	0.0	0.0	6.25	31.25	12.5	6.25	6.25	0.0	0.0
Command text	62.5	12.5	50.0	12.5	25.0	0.0	62.5	0.0	50.0	12.5	25.0	0.0
Claude opus	100.0	0.0	75.0	0.0	56.25	12.5	100.0	0.0	75.0	0.0	56.25	0.0
Claude instant	75.0	6.25	81.25	12.5	43.75	0.0	75.0	0.0	81.25	0.0	43.75	6.25
Claude haiku	100.0	0.0	93.75	0.0	81.25	6.25	100.0	0.0	93.75	0.0	81.25	18.75
Claude sonnet v1	100.0	0.0	87.5	0.0	87.5	6.25	100.0	0.0	87.5	0.0	87.5	0.0
Claude sonnet v2	93.75	6.25	68.75	6.25	25.0	18.75	93.75	6.25	68.75	6.25	25.0	6.25

Table 19: The percentage of correct responses with no redefinition (NR) and the anchored response rate for units of measure redefinitions regarding free-form (FF) responses using ZS prompting.

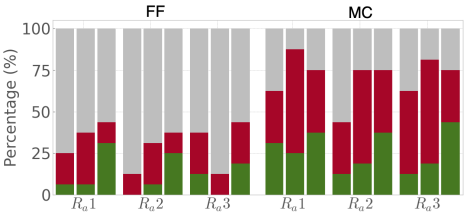
relation between average model performance for all LLMs in the NR case and the number of anchored responses for unit of measure redefinitions using ZS, FS, and CoT prompting, respectively. These correlations mostly exhibit a similar pattern to those observed in the constant redefinition task. However, unlike constants, where high positive correlations were found due to *swapping*, unit of measure redefinitions are implemented using *assignment* exclusively.

Nevertheless, in both ZS and FS prompting, we observe a high positive correlation for Q_3-level questions, similar to the constant redefinition case, denoting increased anchoring for more potent reasoners. This trend holds for the MC response format, but not for the FF format (where correlations are weak), contradicting constants redefinition findings. Apparently, anchoring becomes less prominent with respect to reasoning capability when the LLMs have to generate responses over redefined units of measure.

On the other hand, CoT is evidently capable of reducing anchoring of more potent reasoners, leading to weak or negative correlations in all cases, regardless the redefinition or question difficulty. This is a contradictory fact in comparison to constants redefinitions, revealing that CoT can assist LLMs in reasoning more properly and thus anchor less to their prior knowledge, aligning to basic CoT claims (Kojima et al., 2022).



(a) Response breakdown for Mistral models.



(b) Response breakdown for Llama models.

Figure 12: Results for the different Mistral and Llama models on Q_3 questions using ZS prompting for the redefinition task of units of measure redefinitions. The order of the bars per redefinition type/level corresponds to increasing model size.

Level	R_{a1}	R_{a2}	R_{a3}
	Free-Form (FF)		
Q_1	-0.295	-0.403	-0.33
Q_2	-0.361	-0.247	-0.479
Q_3	-0.063	0.19	0.14
	Multiple Choice (MC)		
Q_1	-0.49	-0.149	-0.542
Q_2	-0.159	-0.023	0.08
Q_3	0.248	0.338	-0.127

Table 20: Correlation between model performance before redefinition with the percentage of anchored answers for each type of unit of measure redefinition and question level in ZS setup. Cells highlighted in pink indicate a **high positive correlation** (> 0.3), while cells in green indicate a **high negative correlation** (< -0.3).

Level	R_{a1}	R_{a2}	R_{a3}
	Free-Form (FF)		
Q_1	-0.32	-0.442	-0.161
Q_2	-0.404	-0.231	0.039
Q_3	0.128	-0.042	0.279
	Multiple Choice (MC)		
Q_1	-0.332	0.058	-0.593
Q_2	0.135	0.131	0.266
Q_3	0.314	0.49	0.101

Table 21: Correlation between model performance before redefinition with the percentage of anchored answers for each type of unit of measure redefinition and question level in FS setup. Cells highlighted in pink indicate a **high positive correlation** (> 0.3), while cells in green indicate a **high negative correlation** (< -0.3).

Level	R_{a1}	R_{a2}	R_{a3}
	Free-Form (FF)		
Q_1	-0.502	-0.598	-0.529
Q_2	-0.465	-0.3	-0.174
Q_3	-0.232	-0.181	-0.079
	Multiple Choice (MC)		
Q_1	-0.528	-0.023	-0.523
Q_2	0.015	-0.091	-0.016
Q_3	-0.127	0.013	-0.242

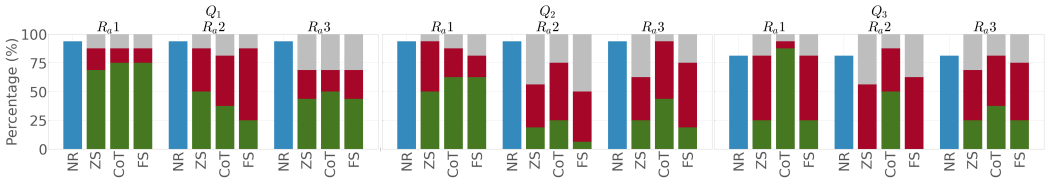
Table 22: Correlation between model performance before redefinition with the percentage of anchored answers for each type of unit of measure redefinition and question level in CoT setup. Cells highlighted in pink indicate a **high positive correlation** (> 0.3), while cells in green indicate a **high negative correlation** (< -0.3).

F Implementation details

We list model cards regarding our employed LLMs in Table 23. All these LLMs are available in Amazon Bedrock[1], a model deployment service provided by Amazon Web Services (AWS), accessed via APIs. The code implementing the API calls, as well as the evaluation part of LLM responses is developed in Kaggle notebooks.

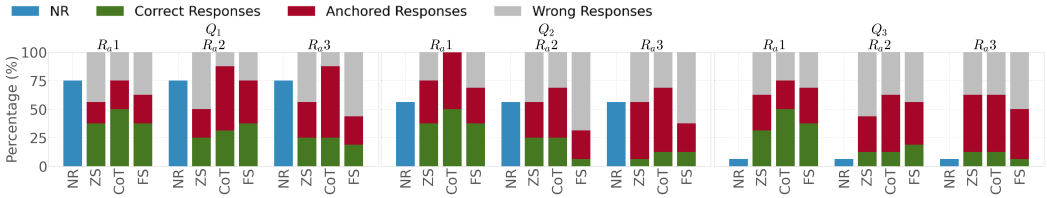
[1] https://aws.amazon.com/bedrock/

(a) Response breakdown for Mistral7B before and after units of measure redefinitions.

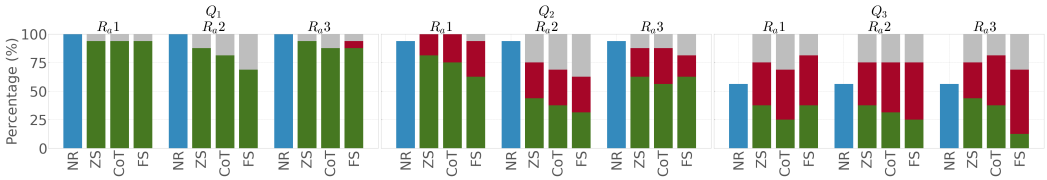


(b) Response breakdown for Mistral Large before and after units of measure redefinitions.

Figure 13: Comparison of Mistral7B and Mistral Large (123B) responses on the MC response format for units of measure redefinitions.



(a) Response breakdown for Llama8B before and after units of measure redefinitions.



(b) Response breakdown for Llama405B before and after units of measure redefinitions.

Figure 14: Comparison of Llama8B and Llama405B responses on the MC response format for units of measure redefinitions.

Finally, all redefinitions are implemented manually by the authors based on engineering textbooks, with the aid of ChatGPT[2] in defining the constants/units to be redefined and as a general guideline towards designing redefinition and question levels.

[2] https://chatgpt.com/

G Prompts

This section illustrates the prompts used for the LLMs. The prompts vary based on the task (NR or redefinition), the required response format (FF or MC), and the prompting techniques selected (ZS, FS, or CoT).

The prompts for the NR task in FF response format are presented below.

prompt_NR_FF_ZS = “““ Answer the following question:
<question>
End the response with the phrase "The final answer is: " followed only by the correct result, with no additional text or commentary. “““

prompt_NR_FF_CoT= “““Answer the following question:
<question>
Let’s think step by step.
End the response with the phrase "The final answer is: " followed only by the correct result, with no additional text or commentary. “““

Model name	Model card	URL
Llama8B	meta-llama/Llama-3.1-8B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Llama70B	meta-llama/Llama-3.1-70B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
Llama405B	meta-llama/Llama-3.1-405B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-405B-Instruct
Mistral7B	mistralai/Mistral-7B-Instruct-v0.2	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
Mixtral8x7B	mistralai/Mixtral-8x7B-v0.1	https://huggingface.co/mistralai/Mixtral-8x7B-v0.1
Mistral Large (123B)	mistral.mistral-large-2402-v1:0	N/A
Claude instant v1	anthropic/claude-instant-1	https://openrouter.ai/anthropic/claude-instant-1
Claude v2	anthropic/claude-2	https://openrouter.ai/anthropic/claude-2
Claude 3 Opus	anthropic/claude-3-opus	https://openrouter.ai/anthropic/claude-3-opus
Claude 3 Haiku	anthropic/claude-3-haiku	https://openrouter.ai/anthropic/claude-3-haiku
Claude 3.5 Sonnet	anthropic/claude-3.5-sonnet	https://openrouter.ai/anthropic/claude-3.5-sonnet
Cohere command light	cohere.command-light-text-v14	N/A
Cohere command text	cohere.command-text-v14	N/A
Cohere command r	CohereForAI/c4ai-command-r-v01	https://huggingface.co/CohereForAI/c4ai-command-r-v01
Cohere command r+	CohereForAI/c4ai-command-r-plus	https://huggingface.co/CohereForAI/c4ai-command-r-plus
Amazon Titan text lite	amazon.titan-text-lite-v1	N/A
Amazon Titan express	amazon.titan-text-express-v1	N/A
Amazon Titan Tg1	amazon.titan-tg1-large	N/A

Table 23: Model cards and hyperlinks for used LLMs. N/A stands for not available hyperlink.

prompt_NR_FF_FS = """Answer the following question:
<question>
Here are some examples of similar questions with their correct answers:
<NR FF examples>
End the response with the phrase "The final answer is: " followed only by the correct result, with no additional text or commentary. """

Below are the prompts for the NR task regarding the MC response format.

---

prompt_NR_MC_ZS = """ Choose A, B, C or D to answer the question:
Question: <question>
A: <A>
B: <B>
C: <C>
D: <D>
Provide only the letter corresponding to the correct answer: "A", "B", "C", or "D".
End the response with the phrase "The final answer is: " followed by the correct letter, with no additional text or commentary. """

prompt_NR_MC_CoT= """Choose A, B, C or D to answer the question:
Question: <question>
A: <A>
B: <B>
C: <C>
D: <D>
Let's think step by step.
Provide only the letter corresponding to the correct answer: "A", "B", "C", or "D".
End the response with the phrase "The final answer is: " followed by the correct letter, with no additional text or commentary. """

prompt_NR_MC_FS = """Choose A, B, C or D to answer the question:
Question: <question>
A: <A>
B: <B>
C: <C>
D: <D>
Here are some examples of similar questions with their correct answers:
<NR MC examples>
Provide only the letter corresponding to the correct answer: "A", "B", "C", or "D".
End the response with the phrase "The final answer is: " followed by the correct letter, with no additional text or commentary. """

For the redefinition tasks, the prompts are identical, with the only difference being the addition of: *"Redefine <X> as <Y>.\n"* at the beginning of each prompt. Additionally, in the few-shot setup, the provided examples included redefined values (constants or units, respectively).

The answers generated by the LLMs were parsed through a different model (Claude 3.5 Sonnet v2) to determine whether they matched the correct response or the anchored one. This approach was necessary because LLMs can produce additional output (e.g. using CoT), making it difficult to extract their answers using regular expressions. The prompts used in this procedure for the NR and redefinition tasks are presented below.

---

prompt_evaluation_NR = """You are tasked with comparing

two answers: one provided by an LLM (the "LLM answer") and the correct answer (the "real answer"). Your job is to determine if the LLM answer matches the real answer.
The comparison should strictly focus on whether the LLM final answer conveys the same meaning or provide the same information as the correct answer. Minor differences in phrasing, wording, or structure are acceptable as long as the core meaning remains identical. For numerical results, differences due to rounding are acceptable as long as the values are reasonably close and within an acceptable margin of error.
Instructions:
1. Compare the LLM answer to the real answer carefully.
2. If the LLM answer matches the real answer, output: correct
3. If the LLM answer does not match the real answer, output: incorrect
Do not output explanations, justifications, or any additional text — only the word "correct" or "incorrect".
LLM Answer:
<llm_answer>
Real Answer:
<real_answer>
Output:
“““

prompt_evaluation_Redefinition = “““You are tasked with comparing an answer provided by an LLM (the "LLM answer") to two reference answers: "Reference Answer 1" and "Reference Answer 2". Your job is to determine if the LLM answer matches either of the two reference answers.
The comparison should strictly focus on whether the LLM final answer conveys the same meaning or provides the same information as one of the reference answers. Minor differences in phrasing, wording, or structure are acceptable as long as the core meaning remains identical. For numerical results, differences due to rounding are acceptable as long as the values are reasonably close and within an acceptable margin of error.
Instructions:
1. Compare the LLM answer carefully with "Reference Answer 1" and "Reference Answer 2".
2. If the LLM answer matches "Reference Answer 1", output: first
3. If the LLM answer matches "Reference Answer 2", output: second
4. If the LLM answer matches neither of the two, output: none
Do not output explanations, justifications, or any additional text — only the words "first", "second", or "none".
LLM Answer:
<llm_answer>
Reference Answer 1:
<reference_answer_1>
Reference Answer 2:
<reference_answer_2>
Output:
“““

Lastly, to categorize the answers as completely wrong, refused, or blank, we prompted Claude 3.5 Sonnet v2 using the following prompt

---

prompt_evaluation_refusal = “““You are tasked with analyzing an LLM answer that does not match either of two reference answers: "Reference Answer 1" and "Reference Answer 2". Your job is to classify the LLM answer into one of the following categories:
1. ***Wrong Answer**: The LLM provided an incorrect response to the question, either factually or logically.
2. ***Blank Answer**: The LLM provided no substantive response, leaving the answer blank or completely empty.
3. ***Refusal to Answer**: The LLM explicitly refused to answer the question, citing reasons such as the question being nonsensical, impossible to answer, or against its guidelines.
Instructions:
1. Analyze the LLM answer and determine which of the three categories it belongs to.
2. If the LLM answer is a **Wrong Answer**, output: wrong
3. If the LLM answer is a **Blank Answer**, output: blank
4. If the LLM answer is a **Refusal to Answer**, output: refusal
5. If the classification is unclear, choose the category that best fits the content of the LLM answer.
Do not output explanations, justifications, or any additional text — only the words "wrong", "blank", or "refusal".
LLM Answer:
<llm_answer>
Reference Answer 1:
<reference_answer_1>
Reference Answer 2:
<reference_answer_2>
Output: “““
