

Does a model’s stated reason for rejecting a candidate do any work?

Archit Rastogi¹

¹*Independent Researcher*

Abstract

Asked to choose between candidates and explain the choice, a language model often rejects a rival by naming a fact its profile lacks: no director, no date of death. That sentence is a claim about the text in front of the model, and it can be tested without any judge. We insert a real corpus sentence stating the named fact into the rival’s profile and ask again under greedy decoding. Two controls separate content from placement: a length-matched irrelevant sentence at the same profile, and the same two sentences at a third option the model never mentioned. In the largest of three runs, six open models on 2WikiMultihopQA, supplying the named fact at the profile the model named moves its choice more than the irrelevant control does, odds ratio 3.57 [1.54, 8.26], Holm $p = 0.0210$, and this survives dropping any single model. The contrast the design was built to detect, the same fact at the option nobody named, does not clear correction, Holm $p = 0.2428$. The strongest result in the family carries no content claim at all: the identical irrelevant sentence moves the choice more at the named rival than at the third option, Holm $p = 0.0008$. Repair and control also differ in co-candidate mentions, relation template and fluency; post-hoc matching on the first two preserves the content effects’ direction, matching fluency weakens one, so the content contrasts bound an effect rather than establish one. A forced single-token probability read disagrees in direction with the free-text choice on that same contrast, and three candidate explanations for the disagreement find no support. Every measurement is a string rule, so each was validated against the records it reads; validation caught eight defects. The largest, a choice-parsing rule that returned the option a model had just rejected in 17.1% of adjudicable responses, would have reported six surviving contrasts instead of four. Code, per-item records with every raw response, and seeded analysis scripts are released.

Keywords

explainability, contrastive explanation, faithfulness, large language models, measurement validity, reproducibility

1. Introduction

When a person explains a decision by ruling an option out, they usually name something it lacks. That sentence carries its own test: supply the missing thing and see whether the decision changes. This is a contrastive explanation, studied in explainable AI before large language models existed [1]. Large language models, asked to pick a candidate and explain the pick, produce the same form: asked why a rival was rejected, they name a concrete fact its profile lacks. Because the named fact is a claim about text the model can see, the test needs no judge. Insert the fact, ask again, and read the choice.

A real item from this study makes the setting concrete (Figure 1). It comes from the six-option design: the question is “Which film has the director died later, The Hole (1960 Film) or Nootrukku Nooru?”, and the model sees six candidates with short profiles: A) The Hole in the Wall (1921 film), B) Thulasi (1987 film), C) The Subterraneans (film), D) Le Masque de la Méduse, E) Nootrukku Nooru, F) Querelle. Llama-3.1-8B-Instruct chooses E and, asked what it ruled out, says “I ruled out option A, The Hole in the Wall, because the profile does not mention the director’s death.” The named defect is a missing date of death. The repair (R1) appends to A’s profile a sentence stating one, taken from sibling D’s profile with the subject swapped: “The Hole in the Wall (1921 film) was Rollin’s final film, as the director died in 2010.” The sentence is not true of A, and need not be: the claim under test is only that the profile omits the attribute (Section 3). The control (R2) appends an unrelated corpus sentence of about the same length to A instead. The same two sentences are appended to C, an option the model never mentioned (R3, R4), and an unedited re-ask (R0) checks that greedy decoding reproduces the original choice. Only R1 moves the choice. Under R1 the model picks A and cites the inserted sentence as its reason; under

LLM4XAI 2026: Workshop on Large Language Models for Explainable AI, co-located with CIKM 2026, November 8, 2026

✉ architrastogi20@gmail.com (A. Rastogi)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

R2 it keeps E and repeats its complaint that A’s profile says nothing of the director’s death; under R3, R4 and R0 it keeps E. On this item the stated reason did work. The forced single-token probability on A, by contrast, barely moves: 0.055 at R0, 0.079 under R1, and 0.089 under R2. The two readouts can disagree on a single item, a pattern RQ3 takes up.

Three questions organise the paper. RQ1, content: does supplying the named fact move the choice more than a length-matched irrelevant sentence at the same profile? This is R1–R2 at the named rival and R3–R4 at the unmentioned option. RQ2, location: does the same sentence move the choice more at the profile the model named than at one it never mentioned? This is R1–R3 for the relevant sentence and R2–R4 for the irrelevant one. RQ3, instrument: do the free-text choice and the probability read agree, and does a judge-free pipeline of string rules measure what it claims to? In the nearest prior tests a model or a learned editor authors the edit [2, 3, 4]; here the experimenter does. The headline numbers come from one run in three designs: Part A, three open 7–8B models at four options; Part B, three further checkpoints at four options; Part C, Part A’s roster at six options. The answers are mixed, and Table 1 states them in one place.

Section 2 places the design among faithfulness tests. Section 3 describes the corpus, the elicited rejection, the five conditions, the integrity gates, and the three runs. Section 4 answers RQ1 and RQ2, including two checks on what the location effect and the date repairs could mean. Section 5 takes up the disagreement between the two readouts. Section 6 reports how the pipeline was validated and what validation caught. Section 7 lists what the design does not yet isolate, and Section 8 closes.

Table 1

What the evidence supports, by research question. Holm p is adjusted over the twelve-test discrete family (Table 3); a contrast is called supported only when it clears that correction. The last column names the check that most qualifies each reading. RQ3 has no single contrast to adjust, so its rows report the checks directly.

Question	Contrast	Holm p	Reading
RQ1: named fact vs. irrelevant control, at the named rival	A R1–R2	0.0210	Supported; survives every single-model exclusion, and grows when neither sentence names a co-candidate (Section 6). Weakens to uncorrected $p=0.099$ on the third of items where repair and control are closest in fluency (Section 7).
RQ1: the same, at the unmentioned option (the designed contrast)	C R1–R2	0.1490	Not supported.
	A R3–R4	0.2428	Not supported.
RQ1: replication on a second roster	C R3–R4	0.0167	Supported, on one model: dropping Mistral takes it to $p=0.30$.
	B R1–R2	1.0000	Not supported.
	B R3–R4	0.2452	Not supported; like B R1–R2, indistinguishable from an underpowered look at Part A.
RQ2: named fact, named vs. unmentioned location	C R1–R3	0.0345	Supported, but fails once rejections that only rank the rival below the choice are dropped (Section 6), and the probability read points the other way.
RQ2: irrelevant sentence, named vs. unmentioned location	C R2–R4	0.0008	Supported; the strongest result, a location effect with no content claim. The rival’s higher starting probability does not explain it in the one testable model, on nine discordant pairs; the tighter band is not significant (Section 4.3).
RQ3: do the two readouts agree?			Same sign on all six content cells; opposite on four of six location cells. Three explanations tested, none supported.
RQ3: does the pipeline measure what it claims?			Validation caught eight defects: three corrected in every number, five quantified, three of those bounded (Section 6). Uncorrected, the parser would report six surviving contrasts, not four.

2. Related work

Contrastive explanation, why this and not that, is how people explain choices, structured around a fact and a foil [1, 5], with NLP-side formulations [6, 7]. ERASER’s faithfulness is automatic, from re-running the model on erased input; only plausibility uses human ratings [8]. A contrastive rejection specifies a falsification condition in the text it produces, reducing that distinction to a string check and a re-ask.

The closest instrument is Atanasova et al.’s counterfactual test of natural language explanations [2]: a learned editor inserts a reason that changes the model’s prediction, and faithfulness is read from whether the explanation reflects that insertion rather than from whether the choice moves. Here the experimenter authors the edit instead, real corpus text with subject substituted, and reads the model’s subsequent choice, with a length-matched content control (a known distractor [9]) and a location control unmatched elsewhere. Wiegrefe et al. instead correlate a model’s label and rationale sensitivity to one perturbation [10]. Ding et al. likewise treat an explanation as a falsifiable hypothesis, generating counterfactual edits for vision-language models by altering a cited concept rather than supplying a missing one [11]. MiCE searches for the minimal edit that flips a classifier to a specified contrast label [12]; here the model states the edit and the experimenter performs it.

Five further lines test a different mechanism. Turpin et al. show chain-of-thought reasoning can omit an influence that steered the answer [13]; Lanham et al. corrupt that reasoning text and check whether the answer changes [14]. Chen et al. have a separate LLM author the edit and humans predict the outcome, finding low precision [3], though a later study reports otherwise [15]. Mayne et al. have the model author its edit and read its size [4]. Matton et al. estimate concept-level causal effects with a Bayesian hierarchical model over LLM-authored counterfactuals [16]. None supplies a fact the model itself named as missing, and in none does the experimenter author the edit from corpus text.

Self-explanation faithfulness varies by explanation type, model, and task [17]; contrastive rejection is another untested case. Explainers disagree on the same model [18]; our disagreement is within one explainer. FaithLM contradicts a stated explanation and rechecks the prediction [19]; supplying a fact the explanation calls missing is the opposite edit. ALCE scores citations by whether the passages cited entail the statement [20], and ContextCite traces which context a generation used [21]; neither edits text to test a stated claim.

3. Method

3.1. Corpus and items

Items come from 2WikiMultihopQA [22]. Each item supplies several candidate entities with short, type-matched profiles: four options in Parts A and B, six in Part C. An attribute missing from one profile is often present, as a real sentence, in a sibling profile in the same item, letting every repair come from the corpus itself, for the attributes it states often enough to source this way: date of birth, date of death, parentage, and director credits.

3.2. Elicitation and the properties of elicited rejections

A model is shown the question and the profiles, picks one, and is asked which candidate it ruled out and what is missing from that candidate’s profile. Asking matters: pooled over three models and the same 40 items, a contrastive rejection appears unprompted in 20.8% of items against 72.5% when asked directly, while accuracy stays flat, 68.4% spontaneous against 68.2% elicited. Asking manufactures the utterance without manufacturing the error, so this paper uses the elicited form throughout. That pilot had set a threshold of 15 usable items per model, and no cell reached it: the best reached 12 of 40.

About 32% of specific rejections name a defect the profile carries, 31.6% spontaneous and 31.8% elicited, a complaint the model’s own input refutes. In 18.8% of attempted items the rejected candidate is the dataset’s correct answer, so we exclude these from every repair arm: repairing the gold answer’s profile is out of scope.

"Which film has the director died later, *The Hole* (1960 Film) or *Nootrukku Nooru*?"
A) *The Hole in the Wall* (1921 film) B) *Thulasi* (1987 film) C) *The Subterraneans* (film) D) *Le Masque de la Méduse* E) *Nootrukku Nooru* F) *Querelle*
model chose E) *Nootrukku Nooru*, and said of A: "I ruled out option A, *The Hole in the Wall*, because the profile does not mention the director's death."

edited profile (each shown once, insertions stacked beneath)	
the named rival (A)	an unnamed third option (C)
A) <i>The Hole in the Wall</i> (1921 film) " (for a 1929 early talkie remake starring Edward G. Robinson and Claudette Colbert see -- <i>The Hole in the Wall</i>.)" <i>The Hole in the Wall</i> is a 1921 American silent drama film produced and distributed by Metro Pictures and directed and co-produced by Maxwell Karger. The film starred Alice Lake and is based on a Broadway play, " <i>The Hole in the Wall</i> " by Fred Jackson.	C) <i>The Subterraneans</i> (film) <i>The Subterraneans</i> is a 1960 film directed by Randal MacDougall based on the 1958 novel of the same name by Jack Kerouac.
R1 (the repair) <i>relevant sentence; re-ask picks A</i> + "The <i>Hole in the Wall</i> (1921 film) was Rollin's final film, as the director died in 2010."	R3 <i>relevant sentence; re-ask picks E</i> + "The <i>Subterraneans</i> (film) was Rollin's final film, as the director died in 2010."
R2 (control) <i>irrelevant sentence; re-ask picks E</i> + "The <i>The Hole in the Wall</i> (1921 film) was the 39th season of the Segunda División de Chile."	R4 <i>irrelevant sentence; re-ask picks E</i> + "The <i>The Subterraneans</i> (film) was the 39th season of the Segunda División de Chile."
R0 <i>nothing edited; re-ask picks E. integrity check: the model must repeat its original choice exactly.</i>	

Figure 1: The five conditions as edited content against edited location, on the real built six-option item walked through in Section 1; the banner lists all six candidates. Columns are locations, the named rival and the third option; each profile shows its two insertions stacked beneath it (R1/R3 relevant, R2/R4 irrelevant length-matched control), with the option the re-ask then picks. R0, below, is the unedited integrity check. Profile text is verbatim from the corpus, punctuation artifacts included; inserted sentences are corpus sentences with the subject substituted; the doubled article in R2 and R4 is one such substitution artifact, found in 3 of Part C’s 563 controls.

3.3. The repair edit and the five conditions

When a rejection names an attribute the profile lacks, the design inserts a real sentence carrying that attribute, drawn from a sibling profile with its subject substituted, into the target profile. The repair need not be true of the target entity: the claim under test is only that the profile omits the named attribute, and any sentence stating it falsifies that claim. Substitution is deterministic: the subject becomes the target entity’s name where that is grammatical, a pronoun where it is not, and sentences where neither works are dropped and counted; that loss compounds across the repair sentence and the independently chosen control, 14.7% and 27.2% of all drops.

Five conditions apply this insertion to different targets and content. R0 re-asks with nothing edited, an integrity check: greedy decoding means the choice must reproduce exactly. Figure 1 shows the remaining four conditions, R1 through R4, as a 2×2 of content against location on the item of Section 1, with R2 and R4 the load-bearing controls. If either moves the choice as often as its relevant counterpart, editing a paragraph disturbs the model regardless of what it says. R1–R2 and R3–R4 isolate the effect of content at fixed location (RQ1); R1–R3 and R2–R4 isolate the effect of location at fixed content (RQ2). Balkir et al. define necessity and sufficiency over removing a feature [23]; these conditions add one instead, to a profile that already lacks it, so neither label transfers unchanged. The primary measure throughout is $P(\text{the model now chooses the edited option})$ against R0, paired per item.

R1 and R3 draw that sentence only from sibling options inside the same item (pool median 1–2). R2 and R4 draw from those siblings plus the entire corpus index, ordered by token-length proximity with no preference for a sibling (pool median 1,430 to 7,241 by part). So in 1,263 of 1,264 built items the control sentence comes from outside the item. Read by whole-word match on the 1,070 items whose sentences rebuild offline, it states a parentage relation in 20–30% of items per part, against 3–4% for the repair.

A switch under a controlled input change is output consistency [24], compatible with mechanisms

Table 2

The eight deterministic integrity gates, from the implementation.

#	Predicate (conditions it applies to)
1	Named attribute absent from the rival’s profile before R1’s edit, present after (R1)
2	R2 edits the rival’s profile only, without introducing the named attribute (R2)
3	R1 and R2’s inserted sentences are within 20% of each other in word count (R1, R2)
4	No inserted date leaves a birth date on or after a death date, in any condition (R1–R4)
5	The inserted sentence contains no other candidate’s full title verbatim, in any condition (R1–R4)
6	The gold answer’s profile is unchanged in every condition (R0–R4)
7	Option order, titles, and question text are identical across all five conditions (R0–R4)
8	R4 edits the same option as R3, without introducing the named attribute (R3, R4)

other than sensitivity to the criterion (Section 7). Work on knowledge conflicts between supplied context and parametric memory bears on this [25, 26]; a substituted sentence checked against the entity’s attested true value has never populated on this corpus, zero of 295, 54 and 125 built items across three option counts, leaving open whether parametric belief enters here.

3.4. Integrity gates

Eight deterministic checks run before any generation, since a failure means the edited passages are wrong; any single failure excludes that R1/R2/R3/R4 combination, and exhausting every combination drops the item. Table 2 enumerates all eight. Gate 8 is the largest single source of exclusion in every part of this design, its selection consequences discussed in Section 7. An early version picked that option blindly and failed on the 65.1% of items already correct at stage one; the reported runs instead prefer, among eligible candidates, whichever already lacks the named attribute, recovering 25–35% more built items per model.

Two of these gates test less than their names suggest; Section 6 quantifies why for both.

3.5. The probability measure and a corrected probe

Alongside the discrete choice, a forced single-token re-ask reads the model’s probability on each candidate letter, giving a continuous Δp per item, exploratory only. The probe places its instruction in the existing final user turn. A first form added a separate user turn, which some chat templates silently accept and one rejects outright; validation caught this during the largest run, and the part already generated was regenerated before any later part began. The form is not cosmetic: on the one model where both were kept, 178 to 207 of 504 non-R0 rows differ by an order of magnitude on the edited-option probability. No row from the first form enters any figure this paper reports.

3.6. Three runs

A first, small run found the choice moving toward a third option at a rate close to the direct repair’s, too small for significance, motivating the matched irrelevant-content control (R4) and this 2×2 design. A second run built that design at scale ($n=295$) but left location open, and its rival’s-profile effect was substantially one model’s: Qwen supplied 77.8% of the pooled net swing from 32% of the items, taking the contrast to $p=0.54$ on exclusion. The third and largest run supplies the headline numbers, with the gate-eight selection preference and the corrected probe above, across three designs: the original three-model roster (Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3) at four options with the corrected probe (Part A, $n=387$ of 1204), three checkpoints untried before, again at four options (Qwen2.5-14B-Instruct-AWQ, Qwen2.5-3B-Instruct, and DeepSeek-R1-Distill-Qwen-14B-AWQ; Part B, $n=314$ of 875), and the original roster at six options (Part C, $n=563$ of 1293). R0 was clean on all nine model-part cells, 0 flips across 1264 rows, so generation is deterministic within a run at this scale.

3.7. Code and data release

Code and per-item records for all three parts are released at <https://github.com/ArchitRastogi20/contrastive-rejection-test> under the MIT licence. The release carries the experiment harness, deterministic scoring rules, analysis scripts, and one JSON record per item per condition including every raw model response. Every reported number in Sections 4 to 6 regenerates from those records via the seeded modules under `harness/`, mapped number by number in the release README, so a rerun reproduces the intervals exactly; the audits that rebuild inserted sentences (Sections 4.4 and 6) also need a local copy of the 2WikiMultihopQA validation split.

4. Results

The population under test throughout this section is items surviving all eight gates, drawn only from elicited rejections, a subpopulation skewed toward stage-one-correct items in every part (Section 7 quantifies the skew). Each item's R0 response is paired against its response under one edited condition, read as a choice flip or a probability change. Every interval is a percentile bootstrap over item-level paired differences, 10,000 resamples at seed 20260822, pooled across models with no cluster term; discrete tests are exact McNemar with a matched-pairs odds ratio and a log-scale Wald interval. Resampling item ids instead of rows as clusters leaves every content-contrast interval unchanged to two decimals.

Built-item yield rose from 32.1% of attempted items at four options to 43.5% at six, consistent with gate eight's logic: more candidates make a third option already lacking the named attribute easier to find. Figure 2 and Table 3 give every number behind Table 1.

4.1. RQ1: the content effect at each location

Figure 2 shows all four contrasts pooled across Parts A, B, and C, on both measures. Four of the twelve discrete contrasts clear Holm correction: A R1–R2, content at the profile the model actually named; C R3–R4, the same content at the third option already lacking the attribute; C R2–R4, matched irrelevant content compared across those same two locations; and C R1–R3, relevant content compared across the named and unnamed locations. A R1–R2 is the most defensible of the four: it survives every single-model exclusion, exact p 0.0118 dropping Llama-3.1-8B-Instruct, 0.0072 dropping Mistral-7B-Instruct-v0.3, 0.0347 dropping Qwen2.5-7B-Instruct. C R2–R4 survives every exclusion the same way, exact p 0.0118, 0.0012, 0.0002 dropping each model in turn. R2 and R4 carry the identical irrelevant sentence, so this measures where an edit lands rather than what it says. C R3–R4 clears Holm ($p=0.0167$) but rests on one model: dropping Mistral-7B-Instruct-v0.3 alone takes it to $p=0.30$.

The contrast this design was built to detect is Part A's discrete R3–R4 contrast, the content effect at the third option. It does not clear Holm correction, $p=0.2428$. Dropping Qwen2.5-7B-Instruct makes the contrast stronger, exact p 0.0347 to 0.0063; dropping Llama-3.1-8B-Instruct ($p=0.1435$) or Mistral-7B-Instruct-v0.3 ($p=0.3323$) is what removes it.

No run of this design establishes a content effect at both locations at once, in either part.

No pooled contrast shows significant between-model heterogeneity: Cochran's $Q=0.81$ ($I^2=0\%$) for A R1–R2, 0.78 (0%) for C R2–R4, 3.02 (33.9%) for A R3–R4, 2.84 (29.6%) for C R3–R4, all $p>0.2$. With three models, Q has little power, so this is weak evidence of homogeneity.

4.2. RQ1 on a second roster: no discrete replication

Part B tested three checkpoints never run before, at four options, to test whether the content effects generalise. Neither survives Holm correction on Part B's discrete measure (Table 3), though R3–R4's own interval excludes an odds ratio of one, 3.25 [1.06, 9.97]: at $n=310$ a twelve-test correction stops it rather than the point estimate. Resampling each part within itself, since a pooled resample would treat a between-roster difference as noise, both roster differences include zero: R1–R2's is +0.035 [−0.004, +0.073], R3–R4's is −0.000 [−0.036, +0.035]. Neither contrast's non-replication can be

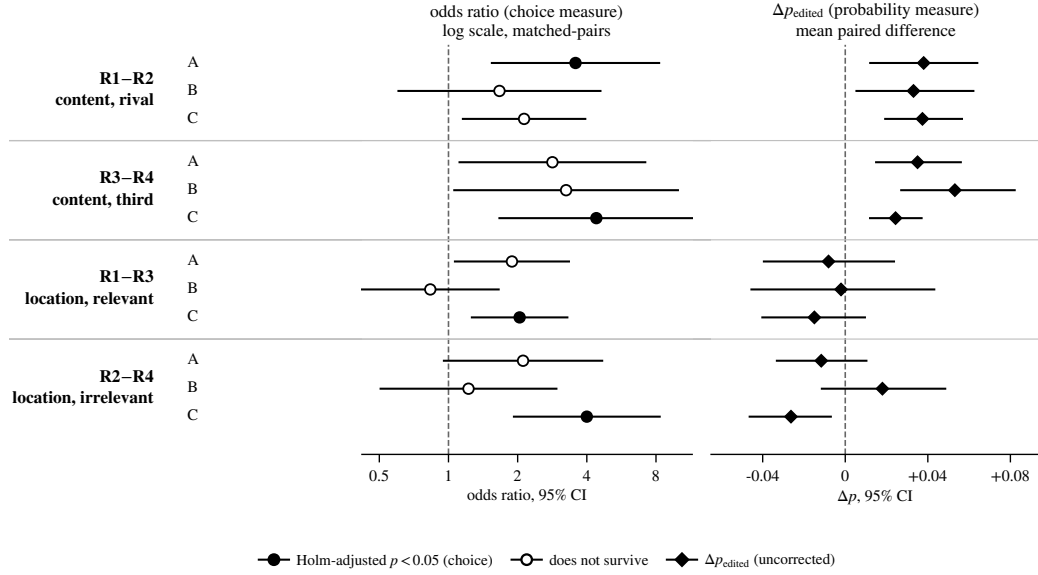


Figure 2: Matched-pairs odds ratios (choice measure, log scale, left) and mean paired Δp_{edited} (probability measure, right), pooled across models, for all four contrasts in all three parts. Solid circles survive Holm correction over the twelve-test discrete family; open circles do not. Diamonds are the continuous measure, outside the Holm family throughout (probe coverage per model is given in Section 5). Table 3 gives the numbers behind every marker.

Table 3

Matched-pairs contrasts behind Figure 2, all four comparisons in all three parts. OR is the matched-pairs odds ratio b/c with an unadjusted Wald interval on the log scale; Holm p is adjusted over the twelve-test discrete family. The two measures do not share a denominator: the discrete read needs a parseable answer (paired n 377–381 in A, 309–310 in B, 547–551 in C) and the continuous read needs a complete letter probe (paired n 299–305 in A, 229 in B, 403–412 in C). DeepSeek-R1-Distill-Qwen-14B-AWQ contributes to every discrete number in Part B and none of the continuous ones, since its probe never completes. Δp is reported at 95% and, like the figure’s diamonds, sits outside the Holm family.

Part	Contrast	b	c	OR [95% CI]	Holm p	Δp [95% CI]
A	R1–R2	25	7	3.57 [1.54, 8.26]	0.0210	+0.0380 [+0.0121, +0.0640]
	R3–R4	17	6	2.83 [1.12, 7.19]	0.2428	+0.0350 [+0.0149, +0.0560]
	R1–R3	34	18	1.89 [1.07, 3.34]	0.2428	−0.0081 [−0.0394, +0.0236]
	R2–R4	19	9	2.11 [0.96, 4.67]	0.3486	−0.0116 [−0.0331, +0.0103]
B	R1–R2	10	6	1.67 [0.61, 4.59]	1.0000	+0.0331 [+0.0054, +0.0621]
	R3–R4	13	4	3.25 [1.06, 9.97]	0.2452	+0.0531 [+0.0271, +0.0821]
	R1–R3	15	18	0.83 [0.42, 1.65]	1.0000	−0.0020 [−0.0454, +0.0431]
	R2–R4	11	9	1.22 [0.51, 2.95]	1.0000	+0.0180 [−0.0113, +0.0484]
C	R1–R2	32	15	2.13 [1.16, 3.94]	0.1490	+0.0374 [+0.0193, +0.0565]
	R3–R4	22	5	4.40 [1.67, 11.62]	0.0167	+0.0244 [+0.0120, +0.0370]
	R1–R3	51	25	2.04 [1.26, 3.29]	0.0345	−0.0150 [−0.0402, +0.0095]
	R2–R4	36	9	4.00 [1.93, 8.30]	0.0008	−0.0263 [−0.0463, −0.0070]

distinguished from an underpowered look at Part A’s own effect. The probability measure, which cannot read the reasoning model (Section 5), moves in the content effect’s direction on the other two (Figure 2). All three Part B checkpoints are Qwen lineage, the family the second run’s analysis flagged as carrying most of the rival’s-profile effect: two Qwen2.5 sizes and a reasoning model distilled onto a Qwen architecture. So no independent lineage beyond Llama and Mistral was tested, and the six-model scope of this paper is, on independent architectures, closer to three. A check on two more independent lineages, Phi-3.5-mini-instruct (3.8B) and Granite-3.0-8B-Instruct, pooled at 186 items, repeats both content contrasts on the continuous measure: R1–R2 +0.0517 [0.0185, 0.0866], R3–R4 +0.1094 [0.0674,

0.1557]. Its discrete measure holds direction on R1–R2 without significance ($p=0.096$) and is significant the other way on R1–R3 ($p=0.001$), opposite its own continuous reading. This repeats the content effect on the continuous measure only; the two-measure disagreement recurs.

4.3. RQ2: the location contrasts

R1–R3 asks whether relevant content moves the choice more at the location a model named than at the one it did not; R2–R4 asks the same question of the irrelevant control. Part C’s R1–R3 clears Holm correction, $p=0.0345$; Part A’s stays non-significant at 0.2428. The continuous measure disagrees: Part C’s R1–R3 Δp is -0.0150 [-0.0402 , $+0.0095$], opposite in sign and including zero (Section 5.1).

C R2–R4 is the opposite case: content is held fixed and only location varies, yet it clears Holm correction most comfortably ($p=0.0008$, Section 4.1) and survives every single-model exclusion. R1–R2 and R3–R4 hold location fixed, so a location artefact cannot manufacture either result. What it does undercut is the comparison between them, since an edit’s effect here depends on where it lands.

Every re-ask is a fresh single-turn prompt, so the model never sees its earlier rejection; the rival is simply the option the model treats as the live alternative. On one reading, the runner-up is where any disturbance is felt first, whatever its probability. On the other, the rival starts higher, at about twice the third option’s unedited probability (Table 4), and needs a smaller push to cross the argmax threshold. The two may be one mechanism; the second is testable by restricting to items where the locations start close. In the band where the two R0 probabilities lie within a factor of three of each other and both exceed 0.01, Part C holds 87 items, and the contrast keeps its direction and its nominal, uncorrected significance, $b=8$ against $c=1$, exact $p=0.039$, risk difference $+0.080$ [$+0.023$, $+0.149$]; Part A’s band holds 74 items, $b=9$ against $c=2$, $p=0.065$. In the tighter band, within a factor of 1.5, neither part is significant, $b=4$ against $c=0$ in C and 7 against 1 in A. The bands are almost entirely one model’s: Llama-3.1-8B-Instruct supplies 85 of Part C’s 87 items and 72 of Part A’s 74, because the other two models put near-zero probability on one of the two locations in nearly every item. So this is a within-Llama check on nine discordant pairs, against the roughly 728 an 80% power calculation would need, and it does not reach Qwen or Mistral, where the pooled effect also lives. It leaves the baseline reading not supported and not ruled out. On this contrast the probability measure points the other way, its interval entirely below zero (Section 5).

4.4. Do date repairs change the right answer?

A date inserted into a profile can change which option the question itself favours, so a switch toward the repaired option might be appropriate use of new evidence rather than sensitivity to the stated reason. Two checks bound this. First, the content contrasts were recomputed inside and outside the stratum of order questions with a date repair, 594 built items. In Part A both are larger outside it, R1–R2 OR 5.33 against 2.25 and R3–R4 5.00 against 1.75; in Part C R3–R4 is 19.0 against 2.60, while R1–R2 is the one reversal, 2.00 against 2.17, with 38 of its 47 discordant pairs inside the stratum. The effects are not confined to items where the date could bear on the answer. Second, for the 1,070 built items that can be rebuilt offline, each inserted year was compared with the competitor’s year in the direction the question asks. Most rows cannot be scored: 352 questions compare a related entity’s date (405 of the stratum’s 594 items ask which film’s director died or was born first, and a death date on a film’s profile may be read as the director’s), 82 name only one option, and the edited option is absent from the question in 20 rows under R1 and 52 under R3. Of the 37 scorable R1 rows, 12 tie, because the only sibling stating the date is the competitor itself; 6 become correct and one of those switched; 18 stay incorrect and one switched; one is unreadable. So of 59 R1 switches inside the stratum, 4 fall on scorable rows and one lands on an option the inserted date makes correct; of 36 R3 switches, 2 are scorable and neither does. This bounds question relevance only on the items that name two people directly. For the director questions, most of the stratum, it stays open, and only a gate that tests the inserted value against the question, not against the profile, would close it.

Table 4

R0 (unedited) probability by target location, pooled across models. Part B excludes the reasoning model, whose probe never completes. Continuous results throughout this paper are complete-case, and probe coverage differs from the choice measure’s coverage (Section 5.2).

Part	rival mean	third-option mean
A	0.145	0.075
B (reasoning model excluded)	0.046	0.032
C	0.140	0.060

5. Measurement

5.1. Two measures and three tested explanations

A first-token probability read and a generated-text answer are known to disagree, mismatching above 60% on one subjective dataset [27], with the text answer the more robust under perturbation [27, 28]. The two measures agree in sign on all six content-contrast cells and diverge on four of six location cells; the reversal on the matched R2–R4 contrast does not follow from a first-token artefact (Explanation 2 below). A binary flip read measures strictly less than a distribution read [29], and continuous metrics, though more diagnostic, are more model-sensitive [30].

Before any edit the rival carries $1.94\times$ the third option’s probability in Part A and $2.36\times$ in Part C, and $1.44\times$ in Part B, whose probabilities run far lower overall (Table 4). The gap is expected: the model names as rejected the option it ranks second. Probability mass over lettered options is independently shaped by position and by a prior over the answer symbol [31, 32], which an argmax inherits. Each item’s option order is fixed by a per-item seeded shuffle, reused unchanged by every condition, so letter assignment cannot vary within a pair. That bias is large where it is free to act: pooled R0 choice rejects a uniform distribution over letters in both Part A and Part C (χ^2 , $p < 0.001$ in both), and one model, Llama-3.1-8B-Instruct, drives most of the pooled deviation, within-model χ^2 129.9 on Part A.

Explanation 1: a ceiling effect. An item where the rival already carries high probability needs only a small push to cross the discrete threshold but has little headroom left, so the two measures could diverge without a real effect difference. Normalising by the available headroom should then shrink the asymmetry; instead it grows. Part C’s pooled R2–R4 gap, a continuous quantity no choice-parsing rule touches, moves from -0.026 , 95% CI $[-0.046, -0.007]$ raw to -0.082 , 95% CI $[-0.139, -0.036]$ once trimmed and renormalised this way, the opposite of what a ceiling predicts, and none of the battery’s three tests supports it.

Explanation 2: the discrete measure reads the wrong quantity. An argmax discards within-ranking movement, governed by margin against the best competitor rather than raw probability, so reading margin might reconcile the two measures. It does not: Part C’s pooled R2–R4 margin sides with the probability read and against the flip counts, Δmargin -0.0400 $[-0.0783, -0.0031]$.

Explanation 3: a few large moves outweigh many small threshold crossings. McNemar’s count is a majority vote over sign changes; a mean can be dominated by a handful of large moves in the opposite direction. In Part C, though, the sign count opposes the flip direction, 181 items moving positive against 226 negative, against the mean. The leave-one-out check in Section 4.1 locates the disagreement instead: C R3–R4’s discrete result depends on Mistral-7B-Instruct-v0.3, dropping it alone takes p to 0.30, while its continuous result, and A R3–R4’s and A R1–R2’s, depend on Llama-3.1-8B-Instruct instead. The two measures are not reading the same subset of the roster.

5.2. Probe portability across architectures

The probability measure depends on a forced single-token decode finding every candidate’s letter in the top of the model’s output distribution, itself presuming the model binds answer symbols to option content [33, 34]. That decode is complete on every row for four of the six models tested,

but on only 55.4% of one model's rows at four options and 32.8% at six, and on zero of 425 rows for a reasoning model. The reasoning model's failure is structural: it always opens its reply with a reasoning block [35], so the single decoded token is always the start of that block, never a letter. The other model's failure appears tied to vocabulary tokenization and candidate count rather than the probe's construction, since the same model's discrete-choice read stays complete throughout. Completion also varies by condition for the degraded model, in Part A and Part C only, reproducible by `harness/audit_probe_missingness.py` in the release. The continuous results are therefore complete-case under a missingness pattern that is not random for that model, and only Part C's R2–R4 has been checked for the resulting bias.

6. Validating the instrument

Every measurement here is a deterministic string rule: the parser that reads a choice, the cues that decide whether a profile states an attribute, the gates that accept an edit. Such rules fail silently, so each was validated against the records it reads: the parser against an independent ground-truth rule, the gates by rebuilding every inserted sentence from the corpus, each cue by reading what it matched. Validation caught eight defects. Three are corrected in every reported number; five need new generation to correct, so they are quantified, and three are bounded.

The three corrected defects are in the choice parser. It returned a rejected option as chosen in 17.1% of adjudicable responses; its letter range stopped at four options while Part C offers six, which left 134 of 155 unreadable Part C rows readable once widened; and it lost a chosen option's trailing parenthetical disambiguator on restatement, 21 of 2,960 adjudicable stage-three rows, 16.2% of the 130 exposed. The worked example of Section 1 shows the first: its R2 response opens “E) Nootrukku Nooru” and goes on to rule out A, and the uncorrected rule read that as a choice of A. The corrected parser has zero disagreements against the ground-truth rule over 10,038 rows and zero regressions, with every test passing. On the uncorrected parser's output the same Holm ladder reports six surviving contrasts instead of four: which conclusions a faithfulness study reaches can turn on a string rule nobody thought to test.

The five that remain are these. Gate three's word-count check (Table 2) cannot fail on a built item by construction, since `check_integrity` tests the metric the search minimises, confirmed on 1,070 rebuilt items; a real-tokenizer check is not possible offline. Gate five's verbatim-title test misses a co-candidate mention at rates tracking both contrasts' treatment axis, 28.0% against 4.3%, since the relevant sentence comes from a sibling profile and the irrelevant one does not. The cue that classifies a named attribute as parentage matches by unbounded substring, on “son” inside “person” or “Johnson”, misclassifying 112 of 127 parentage-attribute items, so a stratum selected by that label is not the relation-matched case it appears to be. In 15.7% of built items the rejection only ranks the rival below the choice rather than asserting absence, and dropping them flips C R1–R3 from clearing Holm correction to failing it. The attribute-presence rule matches its cue as an unbounded substring too, so “studied” followed by a year passes as a date of death, in 6 of the 648 date-repair sentences rebuilt under each of R1 and R3 (0.9%). Two further faults, the probe's first form and blind gate-eight selection (Section 3), were caught before the reported runs, so no reported number carries them, and they are not counted among the eight.

Three of the five bear directly on the content contrasts, and each can be bounded by recomputing R1–R2 and R3–R4 on the rebuilt items where the two sentences of a pair agree on the confound, with `harness/audit_confound_bounds.py`. Where neither sentence names a co-candidate, all four Part A and Part C content effects grow against their rebuilt-item baseline, A R1–R2 from OR 3.14 to 4.75 ($b=19$, $c=4$) and C R3–R4 from 4.40 to 9.00: the confound that tracks the treatment most closely works against the effect. Where repair and control agree on stating a parentage relation, both Part C effects grow and both Part A effects weaken without reversing, R1–R2 to OR 2.33 ($p=0.12$). Dropping the six embedded-cue date repairs moves no count by more than one. On the intersection of all three, all four keep an odds ratio above 4 at uncorrected $p \leq 0.04$; Part B's, on at most twelve discordant pairs, move both ways. These strata are post hoc and uncorrected, and do not match on fluency.

7. Limitations

This work spans one corpus and one language; four relation types, dates of birth and death, parentage, and director credits, account for 98% of built items. Every rejection tested here was elicited (Section 3), and the six models reduce to about three independent lineages (Section 4.2).

A switch after repair admits several mechanisms besides sensitivity to the stated criterion, among them changed salience, inconsistency-induced uncertainty, an edit-detection cue, or a generic preference for relation-bearing text. People confabulate similarly: handed back a face they had not chosen, at most 26% noticed the swap; the rest justified a choice they never made [36], and a re-ask cannot rule this out, being itself a second decision. An edit also moves the input off-distribution [37, 38], here by about 1.2 nats/token in every model tested (pooled R1–R2 -1.114 [$-1.182, -1.045$], $n=925$; R3–R4 -1.181 [$-1.246, -1.115$], $n=931$). In the closest-fluency tercile R1–R2 weakens to OR 2.00, $p=0.0987$, while R3–R4 stays nominally significant, OR 3.00, $p=0.0414$, both uncorrected.

The repair sentence and its control therefore differ on four axes: the named attribute itself, by design; fluency; whether the sentence names a third candidate; and whether it states a parentage relation. Post-hoc matching on the last two preserves the direction of every Part A and Part C content effect (Section 6); matching on fluency weakens the Part A named-rival effect. Until repair and control are matched on fluency within one experiment, R1–R2 and R3–R4 bound an effect rather than establish one.

One comparison is impossible in any run of this design: gate eight tests the post-edit R4 profile for the named attribute, so an option that already carries it can never be a target. A variant with random target selection, run on one model, was underpowered and did not exclude zero. Yet that forbidden case, an option already stating the named fact before any edit, is what would separate a runner-up effect from a general location effect: editing elsewhere, a broader criterion predicts movement toward that option and a location effect predicts none. Building it means replacing gate eight, under one GPU hour at this study’s measured throughput; no run reported here does.

Gate eight also skews selection: built items with a readable stage-one answer were correct in 72.9% against 50.2% of the attempted pool in Part A, 85.6% against 67.6% in B, 48.0% against 42.0% in C.

8. Conclusion

The instrument, and the validation that bounds it (Section 6), is what this paper transfers. On RQ1, supplying the named fact moves the choice at the named profile, surviving every single-model exclusion, the twelve-test correction, and post-hoc matching on co-candidate mention, though not conditioning on fluency (Section 7). The same edit at an option no model mentioned, the contrast the design was built to detect, survives in Part C on one model and not in Part A. On RQ2, Part C’s R1–R3 clears correction ($p = 0.0345$) but fails once rank-only rejections are excluded, and the strongest result carries no content claim at all: an edit’s placement alone moves the choice, so every content result needs checking against its location’s disruptiveness. On RQ3, the two readouts part ways exactly where location enters, and validating the pipeline’s string rules changed which conclusions survive.

Whether the model’s stated reason caused the original decision is still unknown, and so, until the confounds of Section 7 are separated, is whether this design’s own effect is the one that would show it.

Declaration on Generative AI

During the preparation of this work, the author used Claude Sonnet 5 in order to: Paraphrase and reword, and Improve writing style. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication’s content.

References

- [1] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* 267 (2019) 1–38. doi:10.1016/j.artint.2018.07.007.
- [2] P. Atanasova, O.-M. Camburu, C. Lioma, T. Lukasiewicz, J. G. Simonsen, I. Augenstein, Faithfulness tests for natural language explanations, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023): Short Papers*, 2023, pp. 283–294.
- [3] Y. Chen, R. Zhong, N. Ri, C. Zhao, H. He, J. Steinhardt, Z. Yu, K. McKeown, Do models explain themselves? counterfactual simulatability of natural language explanations, in: *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 2024, pp. 7880–7904.
- [4] H. Mayne, R. O. Kearns, Y. Yang, A. M. Bean, E. D. Delaney, C. Russell, A. Mahdi, Llms don’t know their own decision boundaries: The unreliability of self-generated counterfactual explanations, in: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, 2025, pp. 24161–24186.
- [5] T. Miller, Contrastive explanation: A structural-model approach, *The Knowledge Engineering Review* 36 (2021) e14. doi:10.1017/S0269888921000102.
- [6] A. Jacovi, S. Swayamdipta, S. Ravfogel, Y. Elazar, Y. Choi, Y. Goldberg, Contrastive explanations for model interpretability, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, 2021, pp. 1597–1611.
- [7] K. Yin, G. Neubig, Interpreting language models with contrastive explanations, in: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, 2022, pp. 184–198.
- [8] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, Eraser: A benchmark to evaluate rationalized NLP models, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 2020, pp. 4443–4458.
- [9] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, D. Zhou, Large language models can be easily distracted by irrelevant context, in: *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, 2023, pp. 31210–31227.
- [10] S. Wiegrefe, A. Marasović, N. A. Smith, Measuring association between labels and free-text rationales, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, 2021, pp. 10266–10284.
- [11] S. Ding, S. Vasa, A. Ramadwar, Explanation-driven counterfactual testing for faithfulness in vision-language model explanations, in: *NeurIPS 2025 Workshop on Regulatable ML*, 2025. ArXiv:2510.00047.
- [12] A. Ross, A. Marasović, M. Peters, Explaining NLP models via minimal contrastive editing (MiCE), in: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 3840–3852.
- [13] M. Turpin, J. Michael, E. Perez, S. R. Bowman, Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting, in: *Advances in Neural Information Processing Systems* 36 (NeurIPS 2023), 2023.
- [14] T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion, K. Lukošiu̇tė, K. Nguyen, N. Cheng, N. Joseph, N. Schiefer, O. Rausch, R. Larson, S. McCandlish, S. Kundu, S. Kadavath, S. Yang, T. Henighan, T. Maxwell, T. Telleen-Lawton, T. Hume, Z. Hatfield-Dodds, J. Kaplan, J. Brauner, S. R. Bowman, E. Perez, Measuring faithfulness in chain-of-thought reasoning, 2023. ArXiv:2307.13702.
- [15] H. Mayne, J. S. Kang, D. Gould, K. Ramchandran, A. Mahdi, N. Y. Siegel, A positive case for faithfulness: LLM self-explanations help predict model behavior, 2026. ArXiv:2602.02639.
- [16] K. Matton, R. O. Ness, J. Gutttag, E. Kıcıman, Walk the talk? measuring the faithfulness of large language model explanations, in: *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. ArXiv:2504.14150.
- [17] A. Madsen, S. Chandar, S. Reddy, Are self-explanations from large language models faithful?, in: *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 295–337.
- [18] S. Krishna, T. Han, A. Gu, S. Wu, S. Jabbari, H. Lakkaraju, The disagreement problem in explainable

- machine learning: A practitioner’s perspective, *Transactions on Machine Learning Research* (2024). OpenReview jESY2WTZCe; originally posted as arXiv:2202.01602 (2022).
- [19] Y.-N. Chuang, G. Wang, C.-Y. Chang, R. Tang, S. Zhong, F. Yang, A. Wen, M. Du, X. Cai, V. Braverman, X. Hu, FaithLM: Towards faithful explanations for large language models, in: *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026): Long Papers*, 2026, pp. 3802–3824.
 - [20] T. Gao, H. Yen, J. Yu, D. Chen, Enabling large language models to generate text with citations, in: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, 2023, pp. 6465–6488.
 - [21] B. Cohen-Wang, H. Shah, K. Georgiev, A. Mądry, Contextcite: Attributing model generation to context, in: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.
 - [22] X. Ho, A.-K. D. Nguyen, S. Sugawara, A. Aizawa, Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps, in: *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, 2020, pp. 6609–6625.
 - [23] E. Balkir, I. Nejadgholi, K. Fraser, S. Kiritchenko, Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection, in: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2022)*, 2022, pp. 2672–2686.
 - [24] L. Parcalabescu, A. Frank, On measuring faithfulness or self-consistency of natural language explanations, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024): Long Papers*, 2024, pp. 6048–6089.
 - [25] S. Longpre, K. Perisetla, A. Chen, N. Ramesh, C. DuBois, S. Singh, Entity-based knowledge conflicts in question answering, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, 2021, pp. 7052–7063.
 - [26] J. Xie, K. Zhang, J. Chen, R. Lou, Y. Su, Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts, in: *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. OpenReview auKAUJZMO6.
 - [27] X. Wang, B. Ma, C. Hu, L. Weber-Genzel, P. Röttger, F. Kreuter, D. Hovy, B. Plank, “my answer is c”: First-token probabilities do not match text answers in instruction-tuned language models, in: *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 7407–7416. ArXiv:2402.14499.
 - [28] X. Wang, C. Hu, B. Ma, P. Röttger, B. Plank, Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think, in: *First Conference on Language Modeling (COLM 2024)*, 2024. ArXiv:2404.08382.
 - [29] N. Siegel, O.-M. Camburu, N. Heess, M. Perez-Ortiz, The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024): Short Papers*, 2024, pp. 530–546. ArXiv:2404.03189.
 - [30] K. Zaman, S. Srivastava, A causal lens for evaluating faithfulness metrics, in: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, 2025, pp. 29425–29449. ArXiv:2502.18848.
 - [31] C. Zheng, H. Zhou, F. Meng, J. Zhou, M. Huang, Large language models are not robust multiple choice selectors, in: *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. OpenReview shr9PXz7T0.
 - [32] P. Pezeshkpour, E. Hruschka, Large language models sensitivity to the order of options in multiple-choice questions, in: *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 2006–2017.
 - [33] J. Robinson, C. M. Rytting, D. Wingate, Leveraging large language models for multiple choice question answering, in: *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.
 - [34] M. Xue, Z. Hu, L. Liu, K. Liao, S. Li, H. Han, M. Zhao, C. Yin, Strengthened symbol binding makes large language models reliable multiple-choice selectors, in: *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (ACL 2024): Long Papers, 2024, pp. 4331–4344.

- [35] DeepSeek-AI, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. ArXiv:2501.12948; a revised version was later published as “DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning,” *Nature*, vol. 645, pp. 633–638, 2025, doi:10.1038/s41586-025-09422-z.
- [36] P. Johansson, L. Hall, S. Sikström, A. Olsson, Failure to detect mismatches between intention and outcome in a simple decision task, *Science* 310 (2005) 116–119.
- [37] P. Hase, H. Xie, M. Bansal, The out-of-distribution problem in explainability and search methods for feature importance explanations, in: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. ArXiv:2106.00786.
- [38] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, J. Boyd-Graber, Pathologies of neural models make interpretations difficult, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, 2018, pp. 3719–3728.