
BENCHMARKING LLM COMPETENCE ON LOGICAL INFERENCE OVER PROBABILITY OPERATORS

A PREPRINT

Nayera Hasan
Haverford College

Jack Greff
Haverford College

Alvin Grissom II
Haverford College

ABSTRACT

Both expressions of uncertainty and inferences are ubiquitous in natural language, and valid inferences over natural-language expressions of uncertainty are necessary for not only everyday conversations but also for high-stakes domains such as medicine and law. While large language models are increasingly evaluated on logical reasoning tasks, disentangling principled, symbolic reasoning from clever surface-level pattern matching is fraught with difficulty. We introduce a benchmark for reasoning over probability operators—*inference over sentences with gradable epistemic modals* (e.g., *probably*, *might*, *must*) containing 14,320 procedurally-generated English prompts across fifteen inference templates, systematically varying question form, negation strategy, and surface content. Evaluating 29 models, we find that most show answer biases independent of the logical form, a systematic preference for Yes or No. We summarize this with a competence floor: the worse of a model’s accuracy on Yes-correct and No-correct items. Only 9 of 29 models exceed random chance. We also test variations in question form, verb phrases/activity, and both the gender and origin of names used in the prompts, finding biased responses across every axis.

1 Introduction

Epistemic modals (EMs) express a state of knowledge, uncertainty, credence, or belief, and such modals are ubiquitous in human language (Teller, 1972). In English, epistemic modality (EM) is facilitated largely by modal verbs such as *must*, *might*, *may*, and adverbs such as *probably*, and it pervades both everyday and specialist language. The sentences *Bob might be home early* and *Bob will possibly be home early*, for example, express that there is some chance, not necessarily a high chance, that Bob will arrive home early, while the epistemic mood of the sentence *Bob must be home early* expresses certainty that he will be. While modal logics address the subset of cases expressing necessity and possibility—e.g., *must* and *might*—gradable epistemic modality also includes cases expressing *degrees* of certainty, known as *probability operators* (Yalcin, 2010).

In our work, we build upon theoretical work that studies inference rules over natural-language sentences containing probability operators (Yalcin, 2010) to benchmark and analyze the capabilities of LLMs in correctly performing zero-shot logical *inference* with a large set of procedurally-generated EM sentences. EM has received little attention in modern computational linguistics, though earlier work artificial intelligence explored modal and epistemic logics for the integration of uncertain information (Moore, 1981), which is a fundamental problem of AI, broadly defined.

While there exists philosophy and linguistics literature on the semantics of epistemic modals and inference over them (e.g., Yalcin (2010); Kratzer (2012)), there is less work on their analysis in NLP. Holliday, Mandelkern, and Zhang (2024) take an important step in this direction, probing LLMs on modal reasoning with *might* and *must*, representable by modal logic. They find basic errors and logically inconsistent judgments across related inference types, echoing concerns in other areas about the plausible-but-invalid inferences to which LLMs are prone.

This echoes similar work which finds that LLMs’ accuracy on syllogisms depends heavily on recognizing specific token patterns, with changes to names, entities, or quantifiers causing predictable performance shifts (Jiang et al., 2024), which we also examine. EM reasoning competence is particularly relevant given that EM inference is central to nearly all high-stakes, natural-language scenarios in which LLMs—in these contexts, often colloquially called AI—are likely to be deployed: a clinical decision-support system reading a note that a treatment “will probably work” or a

person “probably has a disease” must synthesize premises under uncertainty before ranking options; a system triaging legal documents must distinguish between “might have known” and “probably knew”, since the two carry different evidentiary weight; and a financial pipeline that treats “the market will probably recover” as equivalent to “the market will certainly recover” mischaracterizes risk. As LLMs are embedded in decision loops in medical, legal, and financial domains, their ability to correctly handle inferences with *probably*, *might*, and *must* becomes a practical concern, not only a theoretical one.

Our benchmark covers fifteen templates, encoding thirteen distinct inference patterns (10 valid, 3 invalid) with 14,320 prompts. We vary the wording for logically equivalent inference patterns to check whether the model is changing its inference based on tokens instead of the underlying logic: we use five question forms (*Is it correct that...?*, *Is it true that...?*, *Does it follow that...?*, etc.), multiple negation strategies (prefix, *not*, proposition-level), and surface content (names from seven nationality groups, five activity descriptions). A genuine reasoner will produce the same answer across all of these, irrespective of lexical variation. We evaluate 29 models; most show a fixed bias toward one answer, a systematic preference for *Yes* or *No* that holds across inference types rather than tracking what each inference licenses.

Our primary contributions are: (1) a dataset and benchmark, tested across a variety of models, for inference over probability operators that varies question form, polarity, and content (verb and noun phrases) for fixed logical forms, disentangling models’ answer bias and pattern-matching from principled reasoning (§4); (2) strong evidence that models answer from a fixed answer bias rather than the logical inference (only 9/29 models exceed a random baseline), and that this finding is negation-independent (§4.1); and (3) a comparison of negated sentences and non-negated sentences, revealing accuracy gaps of up to 64 percentage points on semantically identical questions, identifying negation strategy as a confound in benchmark design (§5.1.2).

2 Related Work

To provide theoretical context for the task and situate our work within prior literature on evaluating LLM reasoning, we review related work.

2.1 Modal and Probabilistic Reasoning in LLMs

Prior work on the semantics of epistemic modals provides the theoretical grounding for the inference templates on which we evaluate models. Kratzer (1991, 2012) establishes the standard quantificational framework; Lassiter (2011, 2017) proposes the probabilistic alternative in which *probably* ϕ entails that $p(\phi)$ is high, *might* ϕ that $p(\phi)$ is non-trivial, and *must* ϕ that $p(\phi)$ is near-maximal; and Yalcin (2010)’s work on probability operators catalogues valid and invalid inference patterns involving epistemic comparatives.

Holliday, Mandelkern, and Zhang (2024) test 29 LLMs on 20 inference patterns using conditionals and the modal auxiliaries *might* and *must*. They find that all models commit basic fallacies and that even the strongest model in their study (GPT-4, closed source) exhibits logically inconsistent judgments across related patterns (e.g., accepting Modus Tollens with *must* but rejecting it with *might*, which is logically contradictory). Their templates focus on modal auxiliaries (*might*, *must*) with conditionals, while ours, following semantics work by Yalcin (2010), focuses on the adverb *probably*, examining LLM inference patterns specific to graded epistemic reasoning (e.g., distribution over conjunction, conditional-to-comparative, and conjunction fallacy).

Li, Vrazitulis, and Schlangen (2025) furthermore provide LLMs with short narratives containing factual information and queries the LLMs with prompts containing modal auxiliaries *may/might* vs. *must/have to* as well as with attitude verbs (*know*, *believe*, *doubt*). They note that accuracy is affected by prompt format. Our work is more reductive in its focus on simple templates rather than complex narratives.

Imannezhad, Pothos, and Wills (2026) examine GPT-5’s probabilistic reasoning on conjunction and disjunction fallacies and binary complementarity violations using matched human participant profiles, arguing that GPT-5 appears to have a consistent internal probability model (which our results contravene) that align with quantum-probabilistic models and exceed that of humans.

They elicit numeric probability ratings under a single fixed prompt format and find the judgments internally coherent. In contrast, we hold the inference fixed and vary the question format, finding that binary judgments shift according to the formulation. Our panel is predominantly open-weight for replicability, with a small number of frontier closed models included as reference points; we do not evaluate the commercial GPT-5 model examined by Imannezhad et al., the closest model in our panel being GPT-5.4-mini. Additionally, rather than compare with human judgments, our work focuses on the inherent correctness of LLM inferences.

Macmillan-Scott and Musolesi (2024) tests seven LLMs on cognitive bias tasks from the psychology literature, finding that LLMs are irrational, but that their irrationality does not reflect human patterns: when models give incorrect answers, they are incorrect in different ways than humans. They also found significant inconsistency in responses, consistent with our finding that much of what looks like reasoning on these tasks is surface-driven answer behavior rather than proposition-tracking.

2.2 Surface Sensitivity and Token Bias

LLM performance shifts with surface-form changes that should not affect the answer. Jiang et al. (2024) describe this as “token bias” and introduce a hypothesis-testing framework using McNemar’s test (McNemar, 1947) on matched pairs. They show that on conjunction fallacies and syllogistic problems, changing names, entities, or quantifiers while holding the logic constant produces predictable performance shifts, indicating reliance on superficial patterns rather than reasoning. Similarly, Binz and Schulz (2023) find that GPT-3 (which, being closed source, is no longer available) solved canonical cognitive psychology vignettes “similarly or better than human subjects” but that “small perturbations to vignette-based tasks can lead GPT-3 vastly astray.” Other work highlights that even meaning-preserving *formatting* changes (separator characters, label styles) produce accuracy swings of up to 76 percentage points (Sclar et al., 2024), echoing similar issues documented in machine translation (Shi, Grissom II, and Trinh, 2022). Furthermore, LLMs have a bias toward accepting false presuppositions, making them prone to inferences based on implicit misinformation (Er-makova, Firsov, and Kamps, 2026). Sharma et al. (2024) directly challenges claims of LLM and large reasoning model “reasoning”, demonstrating that the former outperform the latter at low problem complexity but that both collapse at higher problem complexities.

In terms of perturbations, our work—which focuses on a particular, but fundamental, corner of syllogistic reasoning—differs from prior work in several respects. First, perturbations in prior work change what the problem talks about (names, entities, quantifiers) or how the text looks on the page (separators, label styles). We also do this but also change how the question is asked, swapping the metalinguistic wrapper (*Is it true that...*, *Does it follow that...*) and the way a question’s negation is expressed, a dimension of surface sensitivity closer to the logical structure itself. We also decompose accuracy into orthogonal components (the answer bias and polarity sensitivity), going beyond binary detection of sensitivity to measurement of how much each factor explains.

2.3 Response Bias and Acquiescence

Tjuatja et al. (2024) investigate whether LLMs exhibit human-like response biases (acquiescence, response order, opinion floating). Testing nine models on 2,578 question pairs, they find that LLMs generally *do not* reflect human-like bias patterns and that RLHF reduces sensitivity to bias-inducing modifications while increasing sensitivity to non-bias perturbations. Braun (2025) extended this to 37,975 question variations phrased survey-style (“do you agree...”, “don’t you agree...”) over classification tasks on legal-domain texts in English, German, and Polish, finding that LLMs display a bias toward answering *No* in English, the opposite of human acquiescence bias. This aligns with our results. The direction of the answer bias (yes vs. no) is model-dependent rather than universal, with some models strongly yes-biased and others strongly no-biased.

Our baseline takes the minimum of the two answer-conditioned accuracies rather than their mean because the mean alone obscures the bias: mean accuracy can be high when one answer class is near ceiling and the other near floor. The benchmark’s validity-by-negation structure is what makes this minimum informative, providing an absolute baseline under which a maximally biased LLM scores 0 regardless of the valid/invalid mix. We defer the full construction to the discussion of accuracy metrics (§4.1).

A third evaluation framework treats answer-label priors as a calibration problem. Zhao et al. (2021) show that few-shot classification with language models is skewed by majority-label, recency, and common-token biases, estimated the model’s prior from a content-free input; Zhao et al. (2021) divide it out at inference, with large accuracy gains; Fei et al. (2023) refined the prior estimate with in-domain words, and Zhou et al. (2024) with the mean prediction over a test batch. Holtzman et al. (2021) traced related distortions to probability mass splitting across surface forms of the same answer. These results were established on models from 2021 to 2024; our panel suggests the phenomenon has not aged out, with current instruction-tuned and reasoning models still carrying extreme answer priors. Our answer bias is a label bias on a Yes/No space, and this literature suggests such priors are often removable by calibration. Where that work treats the prior as a nuisance to be removed so that accuracy improves, we instead report it directly through the competence floor, which remains diagnostic whether or not the prior could be calibrated away.

2.4 Negation Processing

Truong et al. (2023) evaluate GPT-neo, GPT-3, and InstructGPT on six negation benchmarks, finding three key limitations: insensitivity to the presence of negation, inability to capture its lexical semantics, and failure to reason under negation. They also find that negation shows flat or inverse scaling (larger models are *more* insensitive), a pattern alleviated only by instruction fine-tuning. García-Ferrero et al. (2023) introduced a large negation benchmark of approximately 400,000 sentences, confirming that LLMs rely on superficial cues and that fine-tuning improves but does not fully resolve the problem. So et al. (2025) establish a typology distinguishing morphological negation (*un-*, *in-*) from syntactic negation (*not*), which directly corresponds to the contrast we test. Elkins and Chun (2026) audit negation sensitivity in ethical decisions, finding that open-source models endorse prohibited actions 77% of the time under simple negation and 100% under compound negation.

Our benchmark contains prompts with no negation at all, some expecting Yes and some expecting No, so the core contrast does not depend on negation. On top of that, we implement negation in several ways, asking whether a model that accepts a valid inference (or rejects an invalid one) in the affirmative form can still do so when the question is negated. Implementing it became a source of findings in its own right: writing the same negation as a prefix or with the word *not* produces dramatically different accuracy on semantically identical content (§5.1.2).

3 Benchmark Design

In this section, we describe the overall design of our benchmarks, including the inference templates and their logical forms.

3.1 Inference Templates

Our work is theoretically motivated by prior work on EM semantics which characterizes valid inferences. In particular, we create prompts based on probability operators analyzed by Yalcin (2010), crafted as binary questions with a known correct response.

Thirteen inference templates compose the benchmark of EM expressions; for exposition, we group the templates into three groups based on validity and complexity. Eleven inference rules have a single template; two (distribution over conjunction and conjunctivitis; §5.2) each have two surface-form templates. Altogether, we have fifteen.

Table 1 shows each pattern with a representative prompt from the dataset with the logical form described in terms of probability and modal operators for necessity ($\Box$) and possibility ($\Diamond$). We use $\succeq$ to denote “is at least as likely as”, i.e., $A \succeq B \iff p(A) \geq p(B)$. For brevity, we also use $Pr(A)$ to denote “probably A” and $p(A)$ for a numerical probability, i.e., $Pr(A) \iff p(A) \geq 0.5$ under typical probability assumptions.

3.1.1 Template Descriptions

Simple valid inferences (5 templates) These test straightforward relationships between epistemic operators.

Probably to Might: if ϕ is probable, then ϕ is possible.

Must to Probably: if ϕ is necessary, then ϕ is probable.

Probably to Not Probably Not: if ϕ is probable, then it is not probable that $\neg\phi$.

Positive Form Transfer (PFT) and *Complement Transfer* (CT) test probability preservation under reformulation.

Positive Form Transfer	ψ is at least as likely as ϕ
	probably ϕ .
Complement Transfer	ψ is at least as likely as ϕ
	ϕ is at least as likely as $\neg\phi$.
	probably ψ $\therefore$
	ψ is at least as likely as $\neg\phi$ $\therefore$

Multi-step valid inferences (5 templates) Chancy Modus Ponens and its contrapositive variant Chancy Modus Tollens require combining premises or comparing probabilities.

Chancy Modus Ponens	$\frac{\begin{array}{l} \text{if } \phi \text{ then } \psi \\ \text{probably } \phi. \end{array}}{\text{probably } \psi} \quad \therefore$
Chancy Modus Tollens	$\frac{\begin{array}{l} \text{if } \phi \text{ then } \psi \\ \neg \text{probably } \psi. \end{array}}{\neg \text{probably } \phi} \quad \therefore$

Distribution over conjunction: if probably $(\phi \wedge \psi)$, then probably ϕ and probably ψ .

Conditional to Comparative: if probably $(\phi \rightarrow \psi)$, then ψ is at least as probable as ϕ .

Chancy Disjunction Introduction: if probably ϕ , then probably $(\phi \vee \psi)$.

Invalid inferences (3 templates) These measure correct rejection of fallacious reasoning.

Conjunctivitis (Kyburg Jr, 1970) is a fascinating conjunction fallacy discussed in the philosophical literature. The typical logical assumption of closure under conjunction does not hold for probability operators. Closure under conjunction holds when the truth of ψ and ϕ independently implies the truth of $\psi \wedge \phi$. It holds for non EM predicates but not for the probability operator. I.e., $\text{probably}(\psi), \text{probably}(\phi) \not\Rightarrow \text{probably}(\psi \wedge \phi)$, a fact that can be illustrated by the lottery paradox, whereby we have a collection of propositions of probability at least 0.5 but less than 1, whose conjunction probability necessarily decreases with each new proposition. Since this is a known fallacy discussed in the literature, one might expect that LLMs would be less prone to fall for it.

Might to Probably inference: possibility does not entail probability. $\Diamond\phi \not\Rightarrow \Box\phi$.

Probably to Certain: probability does not entail certainty.

$\text{probably } \phi \not\Rightarrow \Box\phi$.

3.2 Question Form

We present questions in both *affirmative* and *negated* forms to test consistency under negation. In the affirmative condition, we ask whether the conclusion holds (correct answer: *Yes* for valid; *No* for invalid). In the negated condition, we ask whether the same conclusion does not hold (correct answer: *No* for valid, *Yes* for invalid).

Five metalinguistic question forms vary how the question is asked. CORRECT, TRUTH, and VALID negate by switching to the negative adjective (correct to incorrect, true to false, valid to invalid). FOLLOW inserts the word *not* changing, for example, *does it follow* to *does it not follow*). DIRECT negates the proposition itself.

An additional pair of question forms, *Is it not correct that ... ?* and *Is it not valid to conclude that ... ?*, replaces the negative adjective with the word *not* on the same adjective, enabling direct comparison of negation strategies on identical semantic content (§5.1.2). In our analysis, we discuss the perlocutionary force of such questions in light of known issues of LLM sycophancy (§5.1.2). Specifically, for our question templates, introducing negation can pragmatically encourage the model to agree with the questioner. Asking, *Is it not probable that ϕ ?* can, under a common interpretation, sound like the answer should be the *Yes*. We introduce clauses such as *Is it true that* for affirmative and *Is it false that* for the negation to enforce some neutral symmetry in the question surface form without such pragmatic force skewing the results.

3.3 Demographic Bias Controls

Natural language models are known to be prone to spurious biases. To control for this, names are drawn from seven nationality groups (Indian, Russian, Japanese, African, German, French, American) plus abstract letter variables, balanced by gender (common names for men or women). Activities include five semantically similar scenarios. These function as robustness checks: a model that reasons about the logic should show minimal accuracy variation across these controls.

Table 1: Inference templates with formal schema and example prompts (TRUTH question form, affirmative; the negated condition replaces *Is it true that* with *Is it false that* and flips the correct answer). In the affirmative examples shown the correct answer is *Yes* for the valid templates and *No* for the invalid ones, and negation reverses both. $Pr(\phi)$: *probable*; $M(\phi)$: *possible*; $\Box\phi$: *certain*; $\phi \succcurlyeq \psi$: *at least as likely*. Names and activities vary.

Template	Validity	Schema	Example prompt
Probably to Might	Valid	$Pr(\phi) \Rightarrow M(\phi)$	Savir is probably going to be at the party. Is it true that Savir might be at the party?
Must to Probably	Valid	$\Box\phi \Rightarrow Pr(\phi)$	Savir must be at the party. Is it true that it is probable that Savir will be at the party?
Chancy Modus Ponens	Valid	$[\phi \Rightarrow \psi, Pr(\phi)] \Rightarrow Pr(\psi)$	If Savir is going to be at the party, then Ashwin is going to be at the party. Savir is probably going to be at the party. Is it true that Ashwin is probably going to be at the party?
Chancy Modus Tollens	Valid	$Pr(\phi \rightarrow \psi), Pr(\neg\psi) \Rightarrow Pr(\neg\phi)$	If Savir is going to be at the party, then Ashwin is going to be at the party. Ashwin is probably not going to be at the party. Is it true that Savir is probably not going to be at the party?
Distribution over Conjunction	Valid	$Pr(\phi \wedge \psi) \Rightarrow Pr(\phi)$	It is probable that Savir and Ashwin will be at the party. Is it true that it is probable that Savir will be at the party?
Conditional to Comparative	Valid	$Pr(\phi \rightarrow \psi) \Rightarrow \psi \succcurlyeq \phi$	If Savir is going to be at the party, then Ashwin is going to be at the party. Is it true that Ashwin is at least as likely as Savir to be at the party?
Chancy Disjunction Introduction	Valid	$Pr(\phi) \Rightarrow Pr(\phi \vee \psi)$	Savir is probably going to attend the event. Is it true that it is probable that Savir will attend the event or give a presentation?
Positive Form Transfer	Valid	$\phi \succcurlyeq \psi, Pr(\psi) \Rightarrow Pr(\phi)$	Ashwin is at least as likely as Savir to be at the party. Savir is probably going to be at the party. Is it true that Ashwin is probably going to be at the party?
Probably to not probably not	Valid	$Pr(\phi) \Rightarrow \neg Pr(\neg\phi)$	Savir is probably going to be at the party. Is it true that it is not probable that Savir will not be at the party?
Complement Transfer	Valid	$\phi \succcurlyeq \psi, \psi \succcurlyeq \neg\psi \Rightarrow \phi \succcurlyeq \neg\phi$	Ashwin is at least as likely as Savir to be at the party. Savir is at least as likely to be at the party as to not be at the party. Is it true that Ashwin is at least as likely to be at the party as to not be at the party?
Conjunctivitis	Invalid	$Pr(\phi), Pr(\psi) \not\Rightarrow Pr(\phi \wedge \psi)$	It is probable that Savir will be at the party. It is probable that Ashwin will be at the party. Is it true that it is probable that Savir and Ashwin will be at the party?
Might to Probably	Invalid	$M(\phi) \not\Rightarrow Pr(\phi)$	Savir might be at the party. Is it true that it is probable that Savir will be at the party?
Probably to Certain	Invalid	$Pr(\phi) \not\Rightarrow \Box\phi$	Savir is probably going to be at the party. Is it true that it is certain that Savir will be at the party?

3.4 Models Evaluated

We evaluate 29 models in English spanning five size-graded families (Gemma 3: 270M–27B; Gemma 4: E2B–31B; Qwen 3: 0.6B–32B; DeepSeek-R1: 7B–32B; Llama 3: 3B, 8B) plus frontier models (Claude Opus 4.7, Claude Sonnet 4.6, GPT-oss 120B, GPT-oss 20, GPT-5.4-mini), Granite 3.2:8B, Mistral:7b, and the Arabic-oriented Aya-Expanse 8B and 32B.

We show all models the same 80 surface prompts per template, and we run each model once per prompt at temperature $\tau = 0$. All main analyses pool across templates and question forms.

Table 2: The five English question forms and their negated versions, plus the two *not*-prefixed forms, shown on *Must to Probably* (premise: “Savir must be at the party”). The repeated conclusion “it is probable that Savir will be at the party” is abbreviated as Correct answers flip with polarity: Yes affirmative, No negated.

Question form	Affirmative (Yes)	Negated (No)
TRUTH	Is it true that ... ?	Is it false that ... ?
CORRECT	Is it correct that ... ?	Is it incorrect that ... ?
VALID	Is it valid to conclude that ... ?	Is it invalid to conclude that ... ?
FOLLOW	Does it follow that ... ?	Does it not follow that ... ?
DIRECT	Is it probable that Savir will be at the party?	Is it not probable that ...
NOT-CORRECT	—	Is it not correct that ... ?
NOT-VALID	—	Is it not valid to conclude that ... ?

DIRECT carries no metalinguistic wrapper, negating the proposition itself (*probable to not probable*).

Prompting setup We query open-weight models zero-shot, and the first answer token is parsed as *Yes*, *No*, or classified as *Uncertain* otherwise. Two choices bear on interpretation. First, to keep the evaluation strictly zero-shot, we disable chain-of-thought wherever the model permits it: Qwen 3 and DeepSeek-R1 are run with thinking turned off, and GPT-oss is set to its lowest available reasoning level. The reported numbers therefore reflect direct-answer behavior. Second, Qwen 3:4B is queried with a constrained Yes/No output format, because unconstrained generation produced a non-compliance rate as high as 96%. Constraining the output still lets the model choose either answer, so it does not manufacture the answer bias we report.

3.5 Metrics

We score each model on a 2×2 grid, crossing inference validity (valid/invalid) with question negation (affirmative/negated); we refer to its four cells throughout. This crossing determines the correct answer in each cell. We use abbreviations to refer to the accuracy for each intersection of conditions. V.AFF, for example, refers to a *valid* template with a question phrased in the affirmative, while I.NEG refers to an *invalid* template with a negated question. The correct answer is *Yes* in two cells (V.AFF and I.NEG) and *No* in the other two (V.NEG and I.AFF).

Overall accuracy is the fraction of correct responses (with uncertain counted as incorrect), computed over whatever set of prompts a table covers. Since valid templates outnumber invalid ones, weighting the four validity $\times$ negation cells equally instead pulls every strong model toward 0.5, by up to about 9 points; supplementary Table 14 reports accuracy under five weighting schemes so the sensitivity to this choice is visible.

To separate the model’s surface-level bias from principled proposition-tracking, we measure the yes/no bias. We do this by averaging the model accuracy, respectively, on questions for which the answers should be affirmative (*Yes*) and on questions for which the answers should be negative (*No*). For each, we have two cases, the valid and invalid conditions.¹ Let $\text{Acc}(\text{Yes}) = \frac{1}{2}(\text{V.AFF} + \text{I.NEG})$ and $\text{Acc}(\text{No}) = \frac{1}{2}(\text{V.NEG} + \text{I.AFF})$.

The signed gap is the bias toward *Yes* or *Nos*:

$$\text{Bias} = \text{Acc}(\text{Yes}) - \text{Acc}(\text{No}), \quad (1)$$

If a model always guesses *Yes*, the bias will be 1; if it always guesses *No*, it will be -1; and a random baseline has a bias of 0².

We also track polarity sensitivity (PS), the model’s accuracy drop under negation:

$$\text{PS} = \frac{(\text{V.AFF} + \text{I.AFF}) - (\text{V.NEG} + \text{I.NEG})}{2} \quad (2)$$

$$= \frac{[\text{Acc}(\text{non-negated questions}) - \text{Acc}(\text{negated questions})]}{2} \quad (3)$$

Bias is positive when the model favors *Yes*; PS is positive when negation hurts. Bias is an indirect way of measuring a model’s *competence*, i.e., a model’s actual reasoning ability. To measure competence, we take the worse of the two accuracy measurements, the *floor*:

$$\text{Floor} = \min(\text{Acc}(\text{Yes}), \text{Acc}(\text{No})). \quad (4)$$

¹These are true positive rates, but we avoid using the term since “positive” is confusing when discussing yes/no questions that can be negated.

²We have more valid templates than invalid ones; averaging them prevents in this way prevents the valid category from dominating the bias and associated metrics. However, the average still provides useful information.

Intuitively, this related metric concisely describes how poorly the model performs on its worst-performing question polarity (those with a correct answer of either *Yes* or *No*). This is motivated by our observation that models tend to be extremely biased for one answer or the other, irrespective of the internal logic. A constant responder scores 0 on the floor, a coin-flipper 0.5, and a competent reasoner near 1, so the floor has an absolute baseline that overall accuracy lacks.

4 Results

We now describe the EM benchmarking results. While some of the differences in accuracy across conditions may seem small (but many are not), given the query rate of LLMs, seemingly small differences in the number of errors can have huge impacts in the aggregate. If, for example, an LLM is queried 1 billion times, a 0.3% increase in errors is equivalent to 3 million more. Currently, ChatGPT alone receives approximately 2.5 billion queries per day, and GPT-4o received approximately 774 billion queries in 2025, which is still only a fraction of Google’s queries, which now integrate LLMs (Singh, 2026).

4.1 Overall Accuracy

First, we report the aggregate accuracy across all EM prompts for a bird’s eye view of respective model performance as a prelude to more specific analysis. Table 3 presents the four cells for all 29 models, using the metrics defined in §3.5.

Overall accuracy ranges from 20.2% (DeepSeek-R1:7B) to 81.8% (Gemma4:31B), but this obscures the strategies (such as they are) being used. Qwen3:0.6B scores 54.8% overall by nearly always responding *Yes* ($V.AFF = 0.997$, $V.NEG = 0.160$, $I.AFF = 0.030$, $I.NEG = 0.878$), while Llama3.1:8B is balanced (bias only $+0.03$), yet every cell performs near the random baseline, suggesting that the predictions are themselves arbitrary and disconnected from the prompt.

Uncertain responses are rare for most models, below 2%, but high for a handful (e.g., DeepSeek-R1:7b is uncertain 61.5% of the time). Uncertain rates are relatively flat across conditions (expected-Yes 3.2% vs. expected-No 3.4%), so counting them as incorrect does not unduly affect the competence floor. We report rates in Table 3.

Few models clear the floor. Only 9 of 29 models exceed the random Floor baseline of 0.5: Gemma4:31B (0.675), Sonnet 4.6 (0.642), Gemma4:12B (0.614), Qwen3:32B (0.594), Gemma3:27B (0.591), GPT-oss:20 (0.586), Opus 4.7 (0.570), GPT-oss:120B (0.526), and Gemma3:12B (0.518). Even these nine clear it only modestly, by about 2 to 18 points above the 0.5 baseline. Eleven of the 29 models fall below a floor of 0.30. Qwen3:0.6B is the textbook constant Yes responder: near-ceiling where Yes is correct ($V.AFF$ 0.997, $I.NEG$ 0.878), near-zero where No is ($V.NEG$ 0.160, $I.AFF$ 0.030), with 54.8% overall accuracy the average of these extremes.

Competence differences are not only due to negation It is instructive to contrast the competence floor in affirmative vs. negated questions. Recomputing the competence floor from the affirmative cells with $\min(V.AFF, I.AFF)$ achieves this. This affirmative-only floor correlates $r = 0.90$ with the full floor, still only yielding 10 of 29 models above chance, and still leaving the biased models with the worst measures of competence (Qwen3:0.6B at 0.030). Importantly, the answer bias is visible in the affirmative validity contrast alone, so negation is one way to expose it, not necessarily its cause.

5 Analysis

5.1 Robustness to Surface Variation

A genuine reasoner would return the same answer when only the surface of a prompt changes and the logic is held fixed. We examine several such surface variations: the question form and how its negation is written (prefix vs. the word *not*), the activity scenario, and the name’s nationality and gender. The tables here show the models that move most; full per-model tables are in Appendix A.

5.1.1 Question Form

Among affirmative questions, the five question forms land within about four points of each other (67.6–71.4%), a much smaller spread than what we see with negation (29.2–58.4%; Table 4). The collapse is confined to the negated wrapper of the FOLLOW form, *Does it not follow that...?*, which falls to 29.2%, below the 50% chance level and nearly 20 points under the next-lowest wrapper (48.8%). The other four negated wrappers cluster together: Varying adjectives

Table 3: English models, the four cells, sorted by floor; the faded rule separates the nine models that clear the 0.5 floor from those that do not. The four columns V.AFF, V.NEG, I.AFF, I.NEG are per-cell accuracies, so e.g. V.AFF = Acc(V.Aff). **Acc** is overall accuracy. **Uncertain** responses count as incorrect. Bias = Acc(Yes) − Acc(No) (signed; positive = yes-biased). Floor = min(Acc(Yes), Acc(No)) and has a random baseline of 0.5; floors above random chance are in **bold**. Per-condition accuracies below chance are underlined.

Model	Acc	Unc	V.AFF	V.NEG	I.AFF	I.NEG	Bias	PS	Floor
gemma4:31b	0.818	0.0%	0.977	0.761	0.980	<u>0.372</u>	−0.20	+0.41	0.675
Sonnet 4.6	0.751	0.0%	0.994	0.636	0.648	<u>0.491</u>	+0.10	+0.26	0.642
gemma4:12b	0.764	0.0%	0.967	0.754	0.735	<u>0.261</u>	−0.13	+0.34	0.614
qwen3:32b	0.718	0.0%	0.813	0.708	0.827	<u>0.374</u>	−0.17	+0.28	0.594
gemma3:27b	0.648	1.0%	0.913	<u>0.470</u>	0.712	<u>0.334</u>	+0.03	+0.41	0.591
gpt-oss:20	0.737	0.0%	0.815	0.740	0.894	<u>0.356</u>	−0.23	+0.31	0.586
Opus 4.7	0.756	0.0%	0.892	0.762	0.870	<u>0.248</u>	−0.25	+0.38	0.570
gpt-oss:120b	0.724	0.0%	0.727	0.808	0.889	<u>0.326</u>	−0.32	+0.24	0.526
gemma3:12b	0.615	0.0%	0.812	0.620	<u>0.449</u>	<u>0.224</u>	−0.02	+0.21	0.518
gemma4:e4b	0.608	0.0%	0.652	0.655	0.674	0.294	−0.19	+0.19	0.473
llama3.1:8b	0.463	14.3%	<u>0.445</u>	<u>0.435</u>	0.504	0.544	+0.03	−0.01	0.469
gpt-5.4-mini	0.656	0.0%	0.671	0.710	0.887	<u>0.239</u>	−0.34	+0.30	0.455
aya-expanse:32b	0.624	0.1%	0.905	0.522	<u>0.311</u>	0.440	+0.26	+0.13	0.416
gemma3:270m	0.500	0.0%	0.640	<u>0.369</u>	<u>0.456</u>	0.512	+0.16	+0.11	0.413
gemma3:4b	0.529	0.0%	0.603	0.524	<u>0.283</u>	0.589	+0.19	−0.11	0.403
qwen3:8b	0.599	0.0%	0.925	0.410	<u>0.386</u>	0.424	+0.28	+0.24	0.398
qwen3:14b	0.605	0.0%	<u>0.472</u>	0.813	0.729	<u>0.281</u>	−0.39	+0.05	0.377
aya-expanse:8b	0.586	0.6%	0.859	<u>0.467</u>	<u>0.224</u>	0.517	+0.34	+0.05	0.346
mistral:7b	0.535	0.0%	0.836	<u>0.287</u>	<u>0.310</u>	0.604	+0.42	+0.13	0.298
granite3.2:8b	0.607	0.0%	<u>0.406</u>	0.863	0.909	<u>0.169</u>	−0.60	+0.14	0.287
qwen3:4b	0.569	0.0%	<u>0.467</u>	0.699	0.986	<u>0.081</u>	−0.57	+0.34	0.274
llama3.2:3b	0.467	1.5%	<u>0.234</u>	0.605	0.931	<u>0.275</u>	−0.51	+0.14	0.254
gemma4:e2b	0.590	1.7%	<u>0.450</u>	0.797	0.993	<u>0.016</u>	−0.66	+0.32	0.233
deepseek-r1:32b	0.556	2.4%	0.953	<u>0.087</u>	<u>0.349</u>	0.942	+0.73	+0.14	0.218
qwen3:1.7b	0.517	1.5%	0.835	<u>0.286</u>	<u>0.138</u>	0.649	+0.53	+0.02	0.212
gemma3:1b	0.537	0.2%	0.967	<u>0.166</u>	<u>0.101</u>	0.790	+0.74	+0.06	0.133
qwen3:0.6b	0.548	0.0%	0.997	<u>0.160</u>	<u>0.030</u>	0.878	+0.84	−0.01	0.095
deepseek-r1:14b	0.477	10.6%	0.905	<u>0.014</u>	<u>0.134</u>	0.889	+0.82	+0.07	0.074
deepseek-r1:7b	0.202	61.5%	<u>0.415</u>	<u>0.002</u>	<u>0.000</u>	<u>0.357</u>	+0.39	+0.03	0.001

with *Is it false that...?* (58.4%), *Is it incorrect that...?* (56.0%) and *Is it invalid to conclude that...?* (48.8%), together with the proposition-level *Is it not probable that...?* (56.5%), all sit between 49 and 58%. Measured within each form as the affirmative—negated drop (polarity sensitivity), the gap is smallest for the TRUTH form (*Is it true that* vs. *Is it false that*, +0.125) and largest for the FOLLOW form (*Does it follow that* vs. *Does it not follow that*, +0.384); the FOLLOW form is the only one whose negated wrapper leaves the shared range.

The collapse is broad but not uniform: 21 of 29 models score at least 10 points lower on *does it not follow* than on the other negated wrappers, and the strongest models that fall furthest (Gemma4:31B, Sonnet 4.6 and Opus 4.7 each drop 60 or more points), while the handful that do not show it are near-random constant responders already poor in all cases. Tellingly, the DIRECT form’s negated wrapper *Is it not probable that...?* also contains the word *not* yet holds at 56.5%, inside that range, so the difficulty is specific to writing the negation as *not* inside the entailment wrapper (*does it not follow*), not the word *not* itself.

Per model, the same divergence shows up as a wide swing across forms; the nine floor-clearing models have question-form ranges from about 0.19 to 0.50 (Table 5), each concentrated in the *does it not follow* form.

To test whether the negation strategy itself drives this collapse, we added *not*-word versions of the CORRECT and VALID forms (*Is it not correct that...?*, *Is it not valid to conclude that...?*), creating matched pairs where the only difference is whether the negation is written as a prefix (*incorrect/invalid*) or with the word *not*.

5.1.2 Prefix Negation versus *not*-Negation

Since the *not* negation is pragmatically ambiguous—it has a possible interpretation as pining for agreement with a rhetorical question—while prefix negation is unambiguous, these results provide information about the effect of

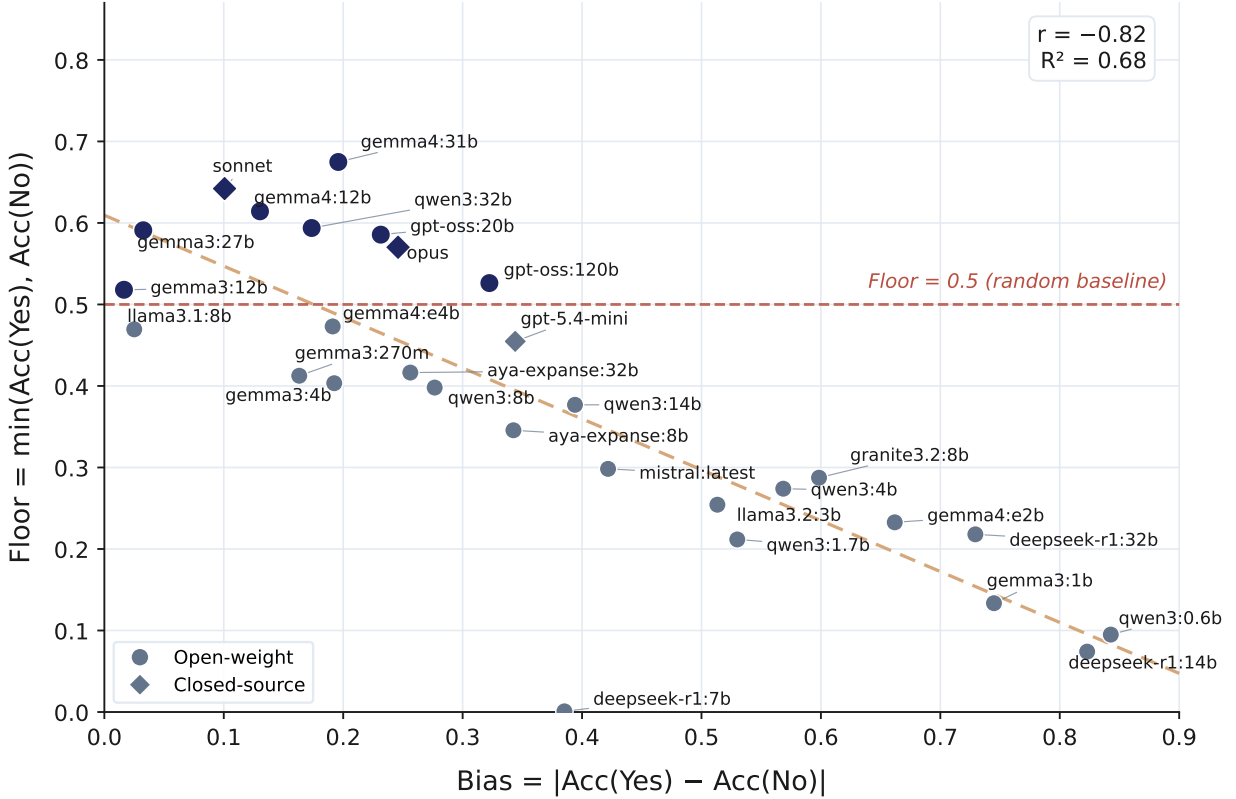


Figure 1: Competence Floor against answer bias ($|\text{Bias}|$) for all 29 models. The two are strongly inversely correlated ($r = -0.82$, $R^2 = 0.68$): the more a model is biased toward one answer, the lower its floor. Only nine models clear the 0.5 random baseline (dashed), and the most biased models (e.g. Qwen3:0.6B) sit near the floor’s zero.

Table 4: Per-question-form results (all English models pooled). **Aff/Neg** are affirmative/negated accuracy (micro-averaged over all templates and models, uncertain counts as incorrect); **PS** is polarity sensitivity (Aff – Neg). Sorted by PS. Under negation only the FOLLOW form’s wrapper, *Does it not follow that ... ?*, collapses to near chance; the other four stay in a tight band, including DIRECT (*Is it not probable that ... ?*), which also uses the word *not*.

Form	Affirmative wrapper	Negated wrapper	Aff	Neg	PS
TRUTH	Is it true that ... ?	Is it false that ... ?	.710	.584	+.125
CORRECT	Is it correct that ... ?	Is it incorrect that ... ?	.699	.560	+.139
DIRECT	Is it probable that ... ?	Is it not probable that ... ?	.714	.565	+.150
VALID	Is it valid to conclude that ... ?	Is it invalid to conclude that ... ?	.676	.488	+.188
FOLLOW	Does it follow that ... ?	Does it not follow that ... ?	.677	.292	+.384

this pragmatic ambiguity versus an unambiguous control. Comparing these matched pairs (prefix *incorrect/invalid* against *not correct/not valid*), the prefix form yields 52.3% accuracy versus 33.6% for the *not* form, an 18.7-point gap, suggesting that the ambiguous form confuses the model. Table 6 presents per-model results. Each model contributes 2,400 paired observations per negation style. The per-model gaps are far larger: Sonnet drops 61.1 points, Opus 59.1, Qwen3:32B 40.3. The direction is not uniform: 25 of 29 models perform better with the prefix form, but 4 (Qwen3:0.6B, Gemma3:1B, Qwen3:1.7B, Llama3.2:3B) show the reverse.

5.1.3 Activity

Replacing the scenario verb phrases (“party”, “event”, “gathering”, “meeting”, or “celebration”) attenuates accuracy. The average per-model range across the five activities is 3.4 pts, relatively minor compared to negation and question-form effects. Only DeepSeek-R1:7b moves substantially (13.2 pts). As with nationality, this is a refusal pattern, its

Table 5: Question-form sensitivity (models that move most). Overall accuracy in each of the five question forms (fraction correct, micro-averaged over all templates and both the affirmative and negated conditions; uncertain counts as incorrect); **Range** is the best–worst spread. The remaining 23 models have Range 0.02–0.32; the smallest belong to committed responders (Qwen3:1.7B 0.02, DeepSeek-R1:7B 0.06). Full per-model table in Table 15.

Model	TRUTH	CORRECT	VALID	FOLLOW	DIRECT	Range
gemma4:31b	0.85	0.87	1.00	0.50	0.88	0.50
Sonnet 4.6	0.82	0.80	0.88	0.50	0.74	0.38
gemma4:12b	0.80	0.83	0.86	0.52	0.82	0.34
gpt-oss:20	0.85	0.84	0.72	0.51	0.77	0.33
Opus 4.7	0.83	0.82	0.84	0.51	0.78	0.33
gpt-oss:120b	0.85	0.79	0.63	0.53	0.82	0.32

Table 6: Negation written as a prefix vs. with the word *not*, for all 29 models. Prefix = *incorrect/invalid* (negation fused into the adjective); *not* = *not correct/not valid* (a separate “not”). Both sides are negated questions with the same proposition and gold answer; only the negation surface form differs. Δ = Prefix accuracy – *not* accuracy. The Yes-rate columns give the model’s percentage of Yes answers under each form.

Model	Accuracy			Yes-rate (%)	
	Prefix	<i>not</i>	Δ	Prefix	<i>not</i>
gemma4:12b	0.776	0.137	+0.639	9.9	75.6
Sonnet 4.6	0.785	0.175	+0.611	23.5	84.5
Opus 4.7	0.781	0.208	+0.572	10.0	66.4
gemma4:31b	0.915	0.372	+0.543	18.1	56.5
qwen3:32b	0.666	0.263	+0.403	37.2	51.2
gpt-5.4-mini	0.632	0.298	+0.334	23.8	58.3
aya-expanse:32b	0.503	0.193	+0.310	37.4	87.5
gpt-oss:20	0.723	0.426	+0.297	32.3	36.0
granite3.2:8b	0.727	0.440	+0.286	10.3	35.8
gpt-oss:120b	0.695	0.446	+0.250	23.1	34.9
gemma3:27b	0.425	0.182	+0.243	53.6	77.5
gemma3:12b	0.502	0.290	+0.212	28.5	66.8
llama3.1:8b	0.510	0.356	+0.154	40.5	55.5
aya-expanse:8b	0.441	0.300	+0.140	64.0	83.4
qwen3:4b	0.531	0.417	+0.113	26.4	35.9
gemma4:e2b	0.679	0.570	+0.108	5.6	21.8
qwen3:8b	0.361	0.255	+0.106	62.8	82.6
qwen3:14b	0.649	0.560	+0.089	30.8	26.4
gemma4:e4b	0.458	0.421	+0.036	38.0	66.1
deepseek-r1:7b	0.103	0.077	+0.026	37.8	32.8
gemma3:270m	0.443	0.428	+0.015	50.9	53.8
deepseek-r1:14b	0.250	0.236	+0.014	93.7	91.1
mistral:7b	0.294	0.284	+0.010	74.2	83.1
gemma3:4b	0.531	0.526	+0.005	50.0	60.3
deepseek-r1:32b	0.261	0.259	+0.002	98.3	98.3
qwen3:0.6b	0.351	0.359	−0.008	82.7	88.2
gemma3:1b	0.245	0.257	−0.012	96.4	93.5
qwen3:1.7b	0.386	0.425	−0.039	79.0	69.5
llama3.2:3b	0.547	0.591	−0.044	31.1	17.1

refusal rate ranging from 46% to 72% across the five activities while accuracy on the items it answers stays near 52%; so the model is sensitive to the verb but expresses that as refusal to answer rather than as different reasoning (Table 7).

5.1.4 Nationality

We find biases are introduced by varying gender, nationality, and activity with every model. As a control, we also use variables X and Y instead of people’s names to compare against. Names are drawn from seven nationality groups, and Table 8 reports each nationality’s accuracy *relative to the abstract variable* (X, Y) baseline. (A positive value means the model does better on that nationality than on the variables.) For 23 of 29 the largest per-nationality gap is at most 4.7 points. The clear exception is DeepSeek-R1:7b, which sits 13 to 34 points below variables on every nationality, being

Table 7: Accuracy by activity scenario (models that move most; uncertain counts as incorrect). **Range** is the spread across the five activities. The remaining 24 models have Range ≤ 0.04 (mean Range = 0.034). Full per-model table in Table 17.

Model	Party	Event	Gather	Meeting	Celebr.	Range
deepseek-r1:7b	0.25	0.28	0.17	0.17	0.15	0.13
qwen3:8b	0.60	0.64	0.58	0.61	0.57	0.06
llama3.1:8b	0.43	0.49	0.49	0.44	0.46	0.06
gpt-oss:20	0.77	0.73	0.72	0.75	0.71	0.06
gemma3:27b	0.65	0.67	0.63	0.66	0.62	0.05

worst on Japanese. It frequently refuses to answer the question, but on the items it actually answers, its accuracy across nationalities spans about 6 points (51–57%), against roughly 21 points on the headline accuracy; the gap comes instead from how often it refuses, and that refusal rate itself swings sharply with nationality, from 46% on Indian names to 87% on Japanese.

Table 8: Accuracy on each nationality group *minus* accuracy on the abstract letter-variable (X, Y) baseline, in percentage points (models that move most; uncertain counts as incorrect). Positive = higher on that nationality than on XY . Ind=Indian, Rus=Russian, Jpn=Japanese, Afr=African, Ger=German, Fre=French, Amr=American. $\max |\Delta|$ is the largest absolute gap. The remaining 23 models have $\max |\Delta| \leq 4.7$ points. Full per-model table in Table 18.

Model	Ind	Rus	Jpn	Afr	Ger	Fre	Amr	$\max \Delta $
deepseek-r1:7b	−13.5	−23.7	−34.1	−20.2	−26.3	−27.8	−23.4	34.1
deepseek-r1:32b	−1.9	−0.5	−1.6	−2.2	−0.8	−9.5	−1.7	9.5
llama3.2:3b	−3.6	−2.2	−9.1	−2.7	−2.8	−8.2	−4.0	9.1
llama3.1:8b	−2.2	+1.9	+3.2	+7.5	+3.8	+4.8	+6.2	7.5
gemma3:12b	−2.1	−3.6	−3.9	−3.8	−2.2	−4.8	−2.8	4.8
Opus 4.7	+2.1	+4.8	+0.9	+1.1	+1.3	+0.0	+1.1	4.8

5.1.5 Gender

Names are balanced by gender, and Table 9 reports each gender’s accuracy relative to the variable (X, Y) control. Women’s and men’s names score within about 3 points of variables for every model except DeepSeek-R1:7b, which exhibits the same refusal behavior seen for nationality. The man-woman name gap itself average 0.6 percentage points; the largest being DeepSeek-R1:32b (+2.3 toward men’s names) and GPT-5.4-mini (+2.2 toward women’s names).

5.2 Conjunction Variants

We test two conjunction inferences, each written two ways that keep the logic fixed (Table 10), with per-model accuracies in Table 11. Recall that distribution over conjunction (DOC) is valid while conjunctivitis is invalid (§3.1). By DOC, $Pr(A \wedge B) \implies Pr(A) \wedge Pr(B)$. To test for consistent reasoning, we vary whether we ask the model about the truth of $Pr(A)$ or $P(B)$, since logically both are true.

On DOC examples, affirmative questions differ by about 5 points on average, and 25 of 29 models stay within 10 points, a large gap for such a simple variation. For conjunctivitis examples how the two premises are worded matters even more. We see this by testing two conditions: one with two people engaging in the same activity and one with two people engaging in different activities. On affirmative items, where the model has to reject the inference, the same-activity wording (two people at one activity) is easier to reject than the dual-activity wording (one person at two activities) by a mean of 12.8 points (up to 96 points for Qwen3:14B, and Sonnet drops from 96.5% to 56.5%). This gap resides entirely in the reject condition. On negated items, where the answer is *Yes*, the two wordings yield very close results (0.3 points), so pooling the two conditions dilutes the effect to 6.6 points, which is why we report it on the affirmative items.

5.3 Per-Template Breakdown

The 13 templates greatly vary in accuracy (42.8% for Conjunctivitis to 74.3% for Probably to Might). The key finding is a dissociation (Table 12): the answer bias varies less across templates than polarity sensitivity does ($|Bias|$ std = 0.056 versus PS std = 0.126, about twice as much), and it stays high on every template (0.42 to 0.62). The bias is thus a model-level property carried across all templates, while negation sensitivity is template-dependent. Model biases also

Table 9: Accuracy on female and male names *minus* accuracy on the variable (X, Y) control, in percentage points (models that move most; uncertain counts as incorrect). Positive = higher than X, Y ; the male–female gap is the difference between the two columns. $\max |\Delta|$ is the larger absolute gap. The remaining 23 models have $\max |\Delta| \leq 3.0$ points. Full per-model table in Table 16.

Model	Female	Male	$\max \Delta $
deepseek-r1:7b	−24.3	−24.0	24.3
llama3.2:3b	−4.8	−4.5	4.8
llama3.1:8b	+4.0	+3.2	4.0
deepseek-r1:32b	−3.8	−1.4	3.8
gemma3:12b	−3.6	−3.0	3.6
qwen3:14b	−2.7	−3.1	3.1

Table 10: The two conjunction inference types, each shown in two surface realizations (affirmative condition; the negated condition swaps *Is it true that* for *Is it false that* and flips the gold answer). The distribution-over-conjunction rows differ only in which name is queried; the conjunctivitis rows differ only in whether the two premises describe two people at one activity or one person at two activities.

Template (logic)	Example (affirmative condition)	Gold
Dist. over conjunction, valid, query first name	It is probable that Savir and Ashwin will be at the party. <i>Is it true that it is probable that Savir will be at the party?</i>	Yes
Dist. over conjunction, valid, query second name	It is probable that Savir and Ashwin will be at the party. <i>Is it true that it is probable that Ashwin will be at the party?</i>	Yes
Conjunctivitis, invalid, same activity	It is probable that Savir will be at the party. It is probable that Ashwin will be at the party. <i>Is it true that it is probable that Savir and Ashwin will be at the party?</i>	No
Conjunctivitis, invalid, dual activity	It is probable that Savir will attend the event. It is probable that Savir will give a presentation. <i>Is it true that it is probable that Savir will attend the event and give a presentation?</i>	No

change direction based on template: on the easiest template (Probably to Might) 28 of 29 models are biased toward *Yes*, so nearly all give the same answer. On the two hardest templates the panel splits roughly in half (Conjunctivitis 13 yes-biased vs. 16 no-biased; Might to Probably 12 vs. 17): the template exerts no shared pull, so which answer a model gives is set mostly by its own bias rather than by the content of the inference.

5.4 Model Scale

Four families provide size variants (Table 13). The answer bias tends to fall as models grow, most clearly in Qwen 3 (Spearman $\rho = -0.89$ between $|\text{Bias}|$ and size) and more weakly in Gemma 3 ($\rho = -0.60$) and Gemma 4 ($\rho = -0.40$), but no family is monotonic (Gemma 3, for instance, spikes at 1B before falling again). DeepSeek-R1 runs the other way, its bias rising with size ($\rho = +0.50$). Thus scaling lowers the bias on average in three of four families, never cleanly, and not at all in the fourth. The FOLLOW question form (the word *not*) shows little scaling improvement across families, suggesting the difficulty with *not* is not resolved by additional parameters.

6 Discussion

Overall, we find widespread biases across multiple dimensions, including name origin, gender, and various perturbations to question formation. More generally, we find definitive evidence that while some models manage to perform above chance, many do not even reach this low bar, and LLMs do not, in general, make basic EM inferences consistently, making them inappropriate for use in question-answering settings that require synthesizing epistemic statements to make valid conclusions.

What accuracy can hide Most models show a fixed answer bias for *Yes* or *No* belied by aggregate accuracy. (Valid reasoning requires both accepting valid inferences and rejecting invalid ones, but we demonstrate conclusively that state of the art models fail to do this. We suggest that future benchmarks using binary questions to evaluate reasoning also

Table 11: Per-model accuracy on the two conjunction templates under a wording change that leaves the logic fixed (affirmative items; uncertain counts as incorrect). **Conjunctivitis** is invalid (correct answer No), so its accuracy is the rejection rate; **Same** states the two premises as two people at one activity and **Dual** as one person at two activities. **Distribution over conjunction** is valid (correct answer Yes); **Name 1** and **Name 2** ask about the first or second name in the shared premise. Δ is the accuracy difference between the two wordings. Rewriting the conjunctivitis premises moves accuracy by a mean of 12.8 points (up to 96), while which name distribution over conjunction asks about moves it by only 4.9 on average. Sorted by the conjunctivitis Δ .

Model	Conjunctivitis (reject rate)			Dist. over conjunction		
	Same	Dual	Δ	Name 1	Name 2	Δ
qwen3:14b	1.00	0.04	+0.96	0.90	0.98	-0.09
qwen3:8b	0.51	0.00	+0.51	1.00	1.00	+0.00
Sonnet 4.6	0.96	0.56	+0.40	1.00	1.00	+0.00
gemma3:12b	0.35	0.01	+0.35	1.00	1.00	+0.00
gemma3:27b	0.70	0.35	+0.35	1.00	1.00	+0.00
gemma4:12b	0.88	0.55	+0.33	0.97	1.00	-0.03
granite3.2:8b	0.96	0.71	+0.26	0.68	0.82	-0.14
qwen3:32b	0.83	0.59	+0.24	0.97	0.99	-0.02
deepseek-r1:32b	0.27	0.04	+0.23	1.00	1.00	-0.00
gpt-oss:20	0.90	0.69	+0.21	0.86	0.84	+0.02
gpt-oss:120b	0.89	0.71	+0.18	0.95	0.93	+0.02
gemma4:e4b	0.49	0.40	+0.09	0.94	0.98	-0.04
gemma3:4b	0.09	0.01	+0.09	0.79	0.95	-0.17
deepseek-r1:14b	0.08	0.00	+0.08	0.99	0.98	+0.01
llama3.1:8b	0.13	0.05	+0.08	0.55	0.64	-0.08
aya-expanse:32b	0.02	0.00	+0.02	1.00	1.00	+0.00
qwen3:1.7b	0.02	0.00	+0.02	1.00	0.99	+0.01
gpt-5.4-mini	0.95	0.94	+0.01	0.23	0.66	-0.43
aya-expanse:8b	0.00	0.00	+0.00	1.00	1.00	+0.00
deepseek-r1:7b	0.00	0.00	+0.00	0.23	0.19	+0.04
gemma3:1b	0.00	0.00	+0.00	1.00	1.00	+0.00
gemma4:31b	1.00	1.00	+0.00	1.00	1.00	+0.00
qwen3:0.6b	0.00	0.00	+0.00	1.00	0.99	+0.01
qwen3:4b	1.00	1.00	+0.00	0.99	0.97	+0.02
Opus 4.7	0.99	1.00	-0.01	0.82	0.95	-0.14
gemma4:e2b	0.98	0.99	-0.01	0.92	0.97	-0.05
llama3.2:3b	0.86	0.89	-0.02	0.43	0.48	-0.05
mistral:7b	0.00	0.06	-0.06	0.99	1.00	-0.01
gemma3:270m	0.07	0.67	-0.60	0.49	0.47	+0.02
Mean	0.52	0.39	+0.13	0.85	0.89	-0.04

report the competence floor, which is cheap to compute and is not inflated by an answer bias. While precision and recall also attempt to address this, they are more difficult to interpret in this specific scenario.

Negation is not the main source of answer bias Two of our probes use negation, and answer biases do not depend on it: the affirmative-only competence floor correlates $r = 0.90$ with the full (affirmative+negation) floor, and the most biased models (Qwen3:0.6B at 0.030) score the worst at both.

What the floor does and does not say The floor measures whether a model performs above chance on both answer classes. The invalid templates are uncontroversial: the conjunction fallacy, possibility not entailing probability, and probability not entailing certainty hold under any reasonable semantics. However, some valid templates rest on stronger commitments. Conditional-to-comparative and probably-to-not-probably-not depend on specific axioms a competent reasoner could contest, though this is not responsible for the results: recomputing the floor with those two templates leaves the number of models with accuracy above 0.5 unchanged at 9 of 29 and moves every above-floor model by at most 0.03 (Sonnet 0.642 to 0.665, Opus 0.570 to 0.567, Qwen3:32B 0.594 to 0.606). Restricting further to a conservative core of only uncontestable inferences (the three invalid templates plus the four most basic valid ones, where *probably* ϕ entails *might* ϕ , *must* ϕ entails *probably* ϕ , distribution over conjunction, and chancy disjunction introduction) tells the same story (10 of 29 above 0.5), and in both restrictions the most biased models remain at the bottom (DeepSeek-R1:7B at 0.00). Biased responders fail the uncontroversial, invalid templates regardless of semantics,

Table 12: Per-template analysis (all English models pooled), grouped by validity. **Acc** is overall accuracy on the template, a micro-average with uncertain counted as incorrect. **|Bias|** is the mean absolute answer bias (from the yes-rate), **PS** the mean polarity sensitivity, and **Agree.** the fraction of models sharing the majority bias direction.

Type	Template	Acc	Bias	PS	Agree.
Valid	Chancy Disjunction Introduction	.574	.539	+.098	62.1%
	Chancy Modus Ponens	.695	.453	+.276	72.4%
	Chancy Modus Tollens	.547	.526	+.095	58.6%
	Complement Transfer	.604	.496	+.183	65.5%
	Conditional to Comparative	.540	.497	+.134	62.1%
	Distribution over Conjunction	.695	.418	+.351	89.7%
	Must to Probably	.633	.509	+.172	62.1%
	Positive Form Transfer	.674	.488	+.023	65.5%
	Probably to Might	.743	.463	+.417	96.6%
	Probably to not probably not	.578	.548	+.335	79.3%
Invalid	Conjunctivitis	.428	.619	+.047	55.2%
	Might to Probably	.447	.608	+.097	58.6%
	Probably to Certain	.683	.462	+.342	79.3%

Table 13: Scaling within size-graded model families (English). **Acc** is overall accuracy with uncertain responses counted as incorrect. **|Bias|** is the magnitude of the signed answer bias and **PS** the polarity sensitivity, both defined in §3.5 (Bias = Acc(Yes) − Acc(No); PS = Acc(affirmative) − Acc(negated), so positive = better on affirmative). **Unc%** is the uncertain rate. Sizes are listed smallest to largest within each family.

Family	Size	Acc	Bias	PS	Unc%
Gemma 3	270M	0.500	0.163	+0.107	0.0
	1B	0.537	0.745	+0.056	0.2
	4B	0.529	0.192	-0.113	0.0
	12B	0.615	0.016	+0.209	0.0
	27B	0.648	0.033	+0.411	1.0
Gemma 4	E2B	0.590	0.662	+0.315	1.7
	E4B	0.608	0.191	+0.189	0.0
	12B	0.764	0.130	+0.343	0.0
	31B	0.818	0.196	+0.412	0.0
Qwen 3	0.6B	0.548	0.843	-0.005	0.0
	1.7B	0.517	0.530	+0.019	1.5
	4B	0.569	0.568	+0.337	0.0
	8B	0.599	0.277	+0.238	0.0
	14B	0.605	0.394	+0.053	0.0
	32B	0.718	0.174	+0.279	0.0
DeepSeek-R1	7B	0.202	0.385	+0.028	61.5
	14B	0.477	0.823	+0.068	10.6
	32B	0.556	0.729	+0.137	2.4

so their low floor cannot be attributed to legitimately contested axioms. An expert or human baseline calibrating how much residual sub-floor behavior is theory disagreement is the natural next step.

Limitations Our evaluation is zero-shot and English-only, and parses the first answer token rather than a free-form justification; whether these patterns persist under longer generation or in other languages is left to future work.

7 Conclusion

We have introduced a benchmark for inference over probability operators that separates a model’s answer bias from proposition-tracking, demonstrating conclusively that models consistently fail at zero-shot reasoning under trivial prompts with an unambiguous correct response. Most models show a fixed bias toward *Yes* or *No*: only 9 of 29 models exceed a competence floor with an absolute random baseline, and the result reproduces from the affirmative validity contrast alone, so it does not appear to depend on negation. Negation variation exposes a strong bias, where prefix and

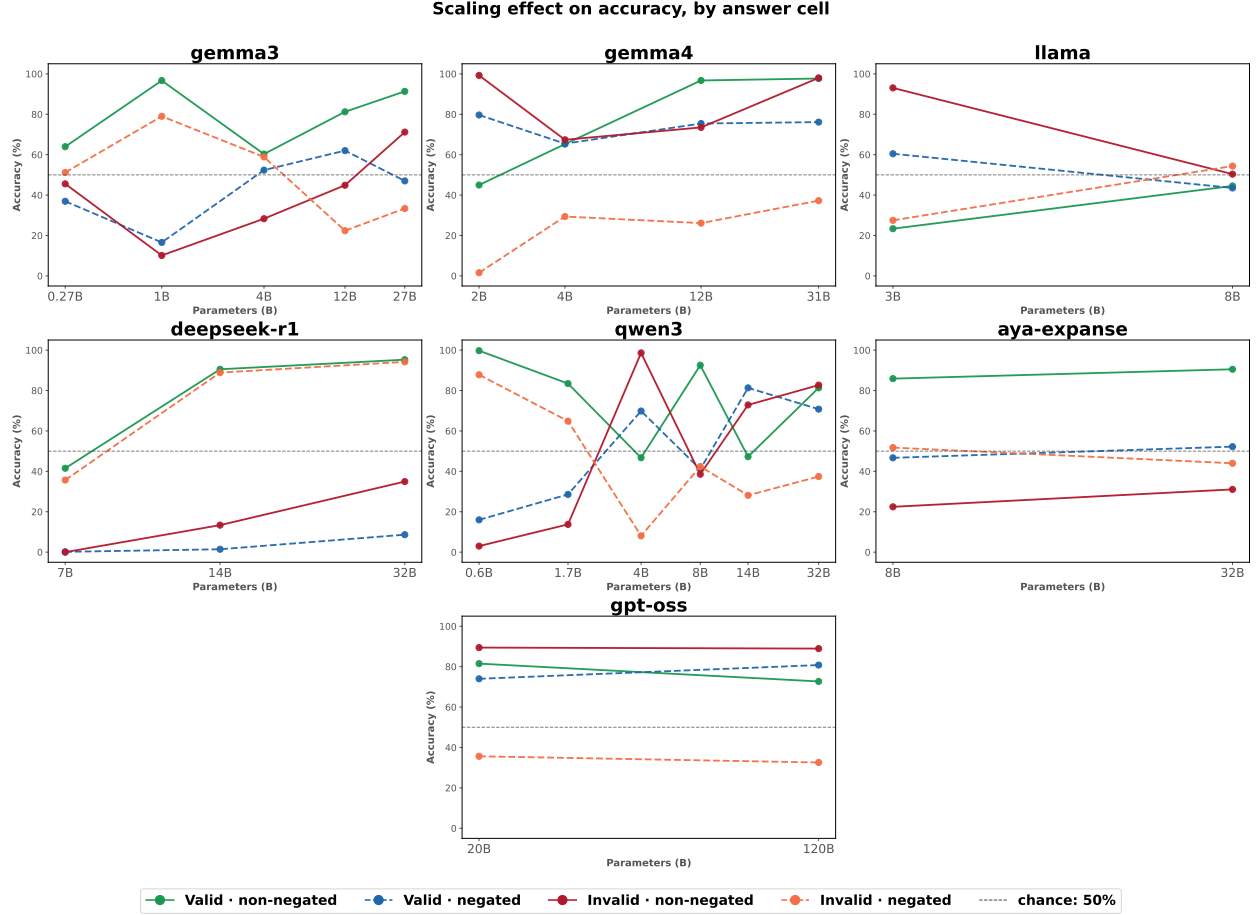


Figure 2: Effect of scaling on accuracy by answer cell (valid/invalid $\times$ negated/non-negated) across model families. Dashed line marks 50% chance. Scaling does not consistently improve accuracy.

not phrasing of the same content differ by up to 64 points. Taken together, these results strongly suggest that part of what is commonly reported as “reasoning” ability may reflect answer bias and is not the result of a coherent internal inference model. A competence floor with an absolute baseline is a useful and easy-to-calculate complement to accuracy as a summary of logical competence. Our prompts are simple and formulaic, and all information necessary to answer correctly is provided in the prompts, suggesting that more subtle ways of expressing the same premises would not lead to successful inferences. Since most inference work limited to English, future work can extend this to other languages.

References

- Binz, Marcel and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.
- Braun, Daniel. 2025. Acquiescence bias in large language models. *arXiv preprint arXiv:2509.08480*.
- Elkins, Katherine and Jon Chun. 2026. When prohibitions become permissions: Auditing negation sensitivity in language models. *arXiv preprint arXiv:2601.21433*.
- Ermakova, Liana, Anton Firsov, and Jaap Kamps. 2026. Confirmation, framing, and position biases in LLM responses. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval (CHIIR '26)*, pages 480–484, ACM.
- Fei, Yu, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of ACL 2023*, pages 14014–14031.

- García-Ferrero, Iker, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. 2023. This is not a dataset: A large negation benchmark to challenge large language models. In *Proceedings of EMNLP 2023*, pages 8596–8615.
- Holliday, Wesley H, Matthew Mandelkern, and Cedegao E Zhang. 2024. Conditional and modal reasoning in large language models. In *Proceedings of EMNLP 2024*, pages 3800–3821.
- Holtzman, Ari, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn’t always right. In *Proceedings of EMNLP 2021*, pages 7038–7051.
- Imannezhad, Pegah, Emmanuel M Pothos, and Andy J Wills. 2026. Divergent patterns of probabilistic reasoning in humans and GPT-5. *Frontiers in Psychology*, 17:1782184.
- Jiang, Bowen, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo J Taylor, and Dan Roth. 2024. A peek into token bias: Large language models are not yet genuine reasoners. In *Proceedings of EMNLP 2024*, pages 4722–4756.
- Kratzer, Angelika. 1991. Modality. In *Semantics: An International Handbook of Contemporary Research*. de Gruyter, pages 639–650.
- Kratzer, Angelika. 2012. *Modals and Conditionals*. Oxford University Press.
- Kyburg Jr, Henry E. 1970. Conjunctivitis. In *Induction, acceptance and rational belief*. Springer, pages 55–82.
- Lassiter, Daniel. 2011. *Measurement and Modality: The Scalar Basis of Modal Semantics*. Ph.D. thesis, New York University.
- Lassiter, Daniel. 2017. *Graded Modality: Qualitative and Quantitative Perspectives*. Oxford University Press.
- Li, Meng, Michael Vrazitulis, and David Schlangen. 2025. Representations of fact, fiction and forecast in large language models: Epistemics and attitudes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27734–27757.
- Macmillan-Scott, Olivia and Mirco Musolesi. 2024. (ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255.
- McNemar, Quinn. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.
- Moore, Robert C. 1981. Reasoning about knowledge and action. In *Readings in artificial intelligence*. Elsevier, pages 473–477.
- Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models’ sensitivity to spurious features in prompt design. In *Proceedings of ICLR 2024*.
- Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Shi, Ruikang, Alvin Grissom II, and Duc Minh Trinh. 2022. Rare but severe neural machine translation errors induced by minimal deletion: An empirical study on Chinese and English. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5175–5180, International Committee on Computational Linguistics, Gyeongju, Republic of Korea.
- Singh, Shubham. 2026. Damensage: Chatgpt statistics (july 2026) – latest active users data. <https://www.damensage.com/chatgpt-statistics/>. [Accessed 13-07-2026].
- So, Yeonkyoung, Gyuseong Lee, Sungmok Jung, Joonhak Lee, JiA Kang, Sangho Kim, and Jaejin Lee. 2025. Thunder-NUBench: A benchmark for LLMs’ sentence-level negation understanding. *arXiv preprint arXiv:2506.14397*.
- Teller, Paul. 1972. Epistemic possibility. *Philosophia*, 2(4):303–320.
- Tjauatja, Lindia, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.
- Truong, Thinh Hung, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: An analysis of language models on negation benchmarks. In *Proceedings of *SEM 2023*, pages 101–114.
- Yalcin, Seth. 2010. Probability operators. *Philosophy Compass*, 5(11):916–937.
- Zhao, Zihao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of ICML 2021*, pages 12697–12706.
- Zhou, Han, Xingchen Wan, Lev Proleeve, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. 2024. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *Proceedings of ICLR 2024*.

A Additional Results

Table 14: Overall accuracy under five pooling schemes (frames *not correct/not valid*/conclude excluded, templates folded to 13, uncertain counts as incorrect). All five score the same responses and differ only in weighting. **micro** is the headline flat mean over every response. **prompt-bal**, **tmpl-bal** and **frame-bal** weight each prompt, template or question form equally. **cat-avg** weights the four validity $\times$ negation cells equally, (V.Aff + V.Neg + I.Aff + I.Neg)/4. Frame-balancing equals micro by construction and prompt-balancing is within a fraction of a point of it; only cat-avg moves the number appreciably, always toward 0.5, because it stops the abundant easy valid cells from dominating the rare invalid-negation cell. **Spread** is the max-min across the five schemes for each model. The mean weighting spread is 4.7 points, and averaged over models cat-avg sits 3.2 points below micro. Sorted by micro.

Model	micro	prompt-bal	tmpl-bal	frame-bal	cat-avg	Spread
gemma4:31b	0.818	0.818	0.829	0.819	0.773	0.056
gemma4:12b	0.764	0.764	0.784	0.764	0.679	0.105
Opus 4.7	0.756	0.782	0.769	0.756	0.693	0.089
Sonnet 4.6	0.751	0.773	0.761	0.750	0.692	0.081
gpt-oss:20	0.737	0.737	0.749	0.737	0.701	0.048
gpt-oss:120b	0.724	0.724	0.733	0.725	0.687	0.046
qwen3:32b	0.718	0.718	0.726	0.718	0.681	0.045
gpt-5.4-mini	0.656	0.656	0.668	0.656	0.627	0.041
gemma3:27b	0.648	0.648	0.665	0.648	0.607	0.058
aya-expanse:32b	0.624	0.624	0.652	0.625	0.545	0.107
gemma3:12b	0.615	0.615	0.631	0.615	0.526	0.105
gemma4:e4b	0.608	0.608	0.613	0.609	0.569	0.044
granite3.2:8b	0.607	0.607	0.606	0.608	0.587	0.021
qwen3:14b	0.605	0.605	0.604	0.605	0.574	0.031
qwen3:8b	0.599	0.599	0.607	0.600	0.536	0.071
gemma4:e2b	0.590	0.590	0.581	0.590	0.564	0.026
aya-expanse:8b	0.586	0.586	0.596	0.586	0.517	0.079
qwen3:4b	0.569	0.569	0.566	0.569	0.558	0.011
deepseek-r1:32b	0.556	0.556	0.559	0.557	0.583	0.027
qwen3:0.6b	0.548	0.548	0.552	0.548	0.516	0.036
gemma3:1b	0.537	0.537	0.535	0.537	0.506	0.031
mistral:7b	0.535	0.535	0.534	0.536	0.509	0.027
gemma3:4b	0.529	0.529	0.532	0.530	0.500	0.032
qwen3:1.7b	0.517	0.517	0.521	0.517	0.477	0.044
gemma3:270m	0.500	0.500	0.499	0.500	0.494	0.006
deepseek-r1:14b	0.477	0.477	0.472	0.476	0.486	0.014
llama3.2:3b	0.467	0.467	0.468	0.467	0.511	0.044
llama3.1:8b	0.463	0.463	0.481	0.464	0.482	0.019
deepseek-r1:7b	0.202	0.202	0.201	0.202	0.193	0.009
Mean	0.597	0.598	0.603	0.597	0.565	0.047

Table 15: Per-model question-form sensitivity, full 29-model table (uncertain counts as incorrect). Overall accuracy in each of the five question forms; **Range** is the best–worst spread. Ordered by competence floor; the faded rule separates the nine models that clear the 0.5 floor from those that do not (§5.1.1).

Model	TRUTH	CORRECT	VALID	FOLLOW	DIRECT	Range
gemma4:31b	0.85	0.87	1.00	0.50	0.88	0.50
Sonnet 4.6	0.82	0.80	0.88	0.50	0.74	0.38
gemma4:12b	0.80	0.83	0.86	0.52	0.82	0.34
qwen3:32b	0.83	0.72	0.72	0.52	0.80	0.31
gemma3:27b	0.73	0.62	0.67	0.53	0.69	0.19
gpt-oss:20	0.85	0.84	0.72	0.51	0.77	0.33
Opus 4.7	0.83	0.82	0.84	0.51	0.78	0.33
gpt-oss:120b	0.85	0.79	0.63	0.53	0.82	0.32
gemma3:12b	0.73	0.66	0.56	0.51	0.62	0.21
gemma4:e4b	0.71	0.68	0.45	0.51	0.70	0.26
llama3.1:8b	0.50	0.58	0.35	0.28	0.61	0.32
gpt-5.4-mini	0.76	0.73	0.64	0.48	0.67	0.28
aya-expanse:32b	0.73	0.69	0.54	0.46	0.71	0.28
gemma3:270m	0.50	0.55	0.47	0.47	0.51	0.08
gemma3:4b	0.57	0.54	0.46	0.44	0.64	0.20
qwen3:8b	0.63	0.62	0.54	0.49	0.72	0.24
qwen3:14b	0.65	0.64	0.51	0.54	0.68	0.17
aya-expanse:8b	0.67	0.62	0.53	0.47	0.64	0.20
mistral:7b	0.61	0.52	0.48	0.43	0.63	0.20
granite3.2:8b	0.68	0.62	0.62	0.45	0.66	0.23
qwen3:4b	0.63	0.60	0.47	0.53	0.61	0.16
llama3.2:3b	0.49	0.51	0.42	0.45	0.46	0.09
gemma4:e2b	0.65	0.65	0.62	0.47	0.56	0.18
deepseek-r1:32b	0.56	0.52	0.55	0.56	0.60	0.07
qwen3:1.7b	0.50	0.52	0.53	0.52	0.51	0.02
gemma3:1b	0.55	0.51	0.50	0.50	0.62	0.12
qwen3:0.6b	0.50	0.51	0.59	0.61	0.54	0.11
deepseek-r1:14b	0.42	0.47	0.52	0.53	0.44	0.11
deepseek-r1:7b	0.17	0.21	0.21	0.23	0.19	0.06

Table 16: Accuracy on female and male names *minus* accuracy on the letter-variable (XY) baseline, all 29 models, in percentage points (uncertain counts as incorrect). Positive = higher than XY. $\max |\Delta|$ is the larger absolute gap. Sorted by $\max |\Delta|$.

Model	Female	Male	$\max \Delta $
deepseek-r1:7b	-24.3	-24.0	24.3
llama3.2:3b	-4.8	-4.5	4.8
llama3.1:8b	+4.0	+3.2	4.0
deepseek-r1:32b	-3.8	-1.4	3.8
gemma3:12b	-3.6	-3.0	3.6
qwen3:14b	-2.7	-3.1	3.1
qwen3:1.7b	+2.9	+3.0	3.0
granite3.2:8b	-2.9	-2.6	2.9
gemma3:27b	-2.8	-1.7	2.8
gpt-5.4-mini	+0.5	-2.7	2.7
gemma4:e2b	-2.4	-2.6	2.6
gemma3:270m	+2.1	+1.8	2.1
gemma3:4b	+1.7	+1.9	1.9
Opus 4.7	+1.3	+1.9	1.9
aya-expanses:32b	-1.6	-1.8	1.8
qwen3:8b	-1.7	-1.3	1.7
qwen3:0.6b	+1.6	+1.7	1.7
deepseek-r1:14b	+1.6	+0.6	1.6
gpt-oss:20	+1.5	+1.0	1.5
gpt-oss:120b	+0.5	+1.1	1.1
gemma4:e4b	+0.1	+1.0	1.0
Sonnet 4.6	-1.0	+0.1	1.0
gemma3:1b	+0.9	+1.0	1.0
qwen3:32b	-0.2	+0.9	0.9
qwen3:4b	+0.2	-0.7	0.7
mistral:7b	+0.6	+0.3	0.6
aya-expanses:8b	+0.6	+0.4	0.6
gemma4:12b	-0.0	-0.5	0.5
gemma4:31b	+0.0	+0.2	0.2

Table 17: Accuracy by activity scenario, all 29 models (uncertain counts as incorrect). **Range** is the spread across the five activities. Sorted by Range.

Model	Party	Event	Gather	Meeting	Celebr.	Range
deepseek-r1:7b	0.25	0.28	0.17	0.17	0.15	0.13
qwen3:8b	0.60	0.64	0.58	0.61	0.57	0.06
llama3.1:8b	0.43	0.49	0.49	0.44	0.46	0.06
gpt-oss:20	0.77	0.73	0.72	0.75	0.71	0.06
gemma3:27b	0.65	0.67	0.63	0.66	0.62	0.05
qwen3:32b	0.70	0.74	0.71	0.73	0.70	0.04
gemma4:e2b	0.58	0.61	0.57	0.61	0.58	0.04
gemma4:e4b	0.63	0.62	0.59	0.59	0.60	0.04
deepseek-r1:32b	0.56	0.58	0.56	0.55	0.53	0.04
llama3.2:3b	0.46	0.48	0.47	0.48	0.45	0.04
mistral:7b	0.56	0.53	0.53	0.54	0.52	0.04
deepseek-r1:14b	0.49	0.49	0.46	0.47	0.46	0.04
granite3.2:8b	0.62	0.61	0.58	0.62	0.61	0.04
qwen3:1.7b	0.51	0.51	0.50	0.54	0.53	0.03
gemma3:12b	0.62	0.62	0.59	0.62	0.62	0.03
gpt-5.4-mini	0.64	0.66	0.66	0.67	0.65	0.03
qwen3:4b	0.56	0.57	0.57	0.58	0.56	0.03
qwen3:0.6b	0.54	0.54	0.55	0.56	0.55	0.03
Opus 4.7	0.75	0.76	0.75	0.77	0.77	0.02
aya-expanse:8b	0.58	0.59	0.57	0.59	0.59	0.02
gemma4:12b	0.76	0.77	0.76	0.77	0.76	0.02
gpt-oss:120b	0.72	0.74	0.72	0.73	0.72	0.02
gemma3:4b	0.52	0.54	0.53	0.53	0.52	0.02
aya-expanse:32b	0.63	0.62	0.62	0.63	0.62	0.02
gemma4:31b	0.82	0.82	0.81	0.81	0.82	0.01
qwen3:14b	0.61	0.61	0.61	0.60	0.60	0.01
Sonnet 4.6	0.75	0.75	0.75	0.76	0.76	0.01
gemma3:270m	0.50	0.50	0.50	0.49	0.50	0.01
gemma3:1b	0.53	0.53	0.53	0.54	0.54	0.01

Table 18: Accuracy on each nationality group *minus* accuracy on the abstract letter-variable (XY) baseline, all 29 models, in percentage points (uncertain counts as incorrect). Positive = higher on that nationality than on XY. Ind=Indian, Rus=Russian, Jpn=Japanese, Afr=African, Ger=German, Fre=French, Amr=American. $\max |\Delta|$ is the largest absolute gap. Sorted by $\max |\Delta|$.

Model	Ind	Rus	Jpn	Afr	Ger	Fre	Amr	$\max \Delta $
deepseek-r1:7b	-13.5	-23.7	-34.1	-20.2	-26.3	-27.8	-23.4	34.1
deepseek-r1:32b	-1.9	-0.5	-1.6	-2.2	-0.8	-9.5	-1.7	9.5
llama3.2:3b	-3.6	-2.2	-9.1	-2.7	-2.8	-8.2	-4.0	9.1
llama3.1:8b	-2.2	+1.9	+3.2	+7.5	+3.8	+4.8	+6.2	7.5
gemma3:12b	-2.1	-3.6	-3.9	-3.8	-2.2	-4.8	-2.8	4.8
Opus 4.7	+2.1	+4.8	+0.9	+1.1	+1.3	+0.0	+1.1	4.8
qwen3:14b	-3.1	-2.8	-2.8	-2.2	-2.1	-4.7	-2.5	4.7
qwen3:1.7b	+2.9	+3.0	+2.4	+4.6	+3.1	+3.2	+1.4	4.6
gemma3:4b	+4.6	+4.1	+0.5	+0.2	+0.9	+0.5	+1.7	4.6
gemma3:27b	-1.1	-0.9	-4.1	-3.3	-1.4	-2.9	-2.1	4.1
deepseek-r1:14b	+2.5	+3.6	-2.8	+3.9	+0.5	-2.2	+2.1	3.9
granite3.2:8b	-3.2	-1.1	-3.2	-3.8	-3.5	-2.0	-2.2	3.8
gemma4:e2b	-2.6	-2.4	-2.9	-3.5	-2.2	-1.6	-2.5	3.5
qwen3:0.6b	+0.3	+3.3	+2.5	-0.5	+1.5	+2.1	+2.2	3.3
gpt-oss:20	+1.4	+0.5	+0.1	+1.9	+1.5	+0.3	+3.3	3.3
gpt-5.4-mini	-1.6	-1.9	-0.9	-0.4	-2.2	-3.2	+2.4	3.2
gemma3:270m	+1.4	+1.0	+3.0	+1.3	+2.0	+2.4	+2.2	3.0
qwen3:8b	-1.6	-0.3	-1.4	-3.0	-1.6	-1.3	-1.4	3.0
gpt-oss:120b	-0.2	+2.2	-0.9	-0.9	+0.8	+1.9	+2.8	2.8
aya-expanse:32b	-1.8	-2.5	-2.1	-0.9	-1.8	-1.3	-1.5	2.5
gemma3:1b	+0.6	+0.7	+2.1	+0.9	+1.3	+0.3	+0.7	2.1
qwen3:32b	-0.5	+0.7	+1.5	+0.3	+1.7	-0.6	-0.7	1.7
mistral:7b	-0.6	+1.7	+1.7	+0.3	+1.1	+0.1	-1.1	1.7
aya-expanse:8b	+0.4	+1.4	+0.4	+0.5	+1.7	-0.8	+0.2	1.7
qwen3:4b	+0.2	-1.5	+0.4	-0.2	+0.3	-0.6	-0.3	1.5
gemma4:12b	+0.9	+0.0	-0.7	+0.9	-1.2	-1.4	-0.5	1.4
gemma4:e4b	+0.6	+1.2	+0.7	+1.2	-0.5	-0.4	+1.2	1.2
Sonnet 4.6	-0.1	+0.6	-0.7	-0.6	-1.1	-0.6	-0.2	1.1
gemma4:31b	+0.4	-0.2	-0.3	-0.1	+0.5	-0.2	+0.5	0.5

