Using Large Language Models to Estimate Belief Strength in Reasoning

Jérémie Beucler[a, *], Zoe Purcell[a], Lucie Charles[b], and Wim De Neys[a]

[a] LaPsyDÉ, CNRS, Université Paris-Cité, F-75005 Paris, France;

[b]School of Biological and Behavioural Sciences, Queen Mary, University of London, London, United Kingdom;

* Corresponding author at: LaPsyDÉ, CNRS, Université Paris-Cité, 46, rue Saint-Jacques, F-75005 Paris, France. E-mail address: jeremie.beucler@gmail.com (J. Beucler).

Preprint accepted for publication in Behavior Research Methods.

Published version: https://doi.org/10.3758/s13428-025-02923-9

Abstract

Accurately quantifying belief strength in heuristics-and-biases tasks is crucial yet methodologically challenging. In this paper, we introduce an automated method leveraging large language models (LLMs) to systematically measure and manipulate belief strength. We specifically tested this method in the widely used "lawyer-engineer" base-rate neglect task, in which stereotypical descriptions (e.g., someone enjoying mathematical puzzles) conflict with normative base-rate information (e.g., engineers represent a very small percentage of the sample). Using this approach, we created an open-access database containing over 100,000 unique items systematically varying in stereotype-driven belief strength. Validation studies demonstrate that our LLM-derived belief strength measure correlates strongly with human typicality ratings and robustly predicts human choices in a base-rate neglect task. Additionally, our method revealed substantial and previously unnoticed variability in stereotype-driven belief strength in popular base-rate items from existing research, underlining the need to control for this in future studies. We further highlight methodological improvements achievable by refining the LLM prompt, as well as ways to enhance cross-cultural validity. The database presented here serves as a powerful resource for researchers, facilitating rigorous, replicable, and theoretically precise experimental designs, as well as enabling advancements in cognitive and computational modeling of reasoning. To support its use, we provide the R package *baserater*, which allows researchers to access the database to apply or adapt the method to their own research.

Keywords: base-rate neglect; belief strength; heuristic; large language models; open-access database.

Using Large Language Models to Estimate Belief Strength in Reasoning

Consider the following problem adapted from the classic "lawyer-engineer" problem (Kahneman & Tversky, 1973):

> There is an event with 1,000 people, of which 996 are lawyers and 4 are engineers. Jack is a randomly chosen participant who attended the event. He is a 45-year-old man, who is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles. Which is more likely: that Jack is an engineer or a lawyer?

Despite the overwhelming base-rate probability that only 4 out of 1,000 attendees are engineers, most people incorrectly choose the "engineer" option—reacting to stereotypical cues, such as an affinity for mathematical puzzles, rather than the base-rate information. This inclination to overlook base-rate information is known as base-rate neglect (for a review, see Pennycook et al., 2022), a well-documented cognitive bias with significant implications for decision-making in various domains, including medicine (e.g., Bergus et al., 1995) and justice (Thompson & Schumann, 2017).

Base-rate neglect and similar cognitive biases are well accounted for by dual-process theories, which describe human thinking as an interplay between fast, effortless intuitive ("System 1") processing and slower, more effortful, deliberative ("System 2") processing (e.g., Evans & Stanovich, 2013; Kahneman, 2011). According to these theories, cognitive biases arise primarily because individuals frequently rely on System 1 processing, using heuristic cues—mental shortcuts—that facilitate quick but often inaccurate decisions.

Although dual-process theories have significantly advanced our understanding of human cognition, precisely how heuristic and logical-probabilistic information (e.g., base-rates) interact

within these systems remains unclear. One critical barrier to resolving this issue has been methodological: researchers lack precise, systematic methods to measure and manipulate belief strength—the degree to which heuristic cues influence reasoning (also sometimes referred to as heuristic strength in the literature). A belief corresponds to prior knowledge about the world. In the reasoning literature, the role of belief has been most clearly articulated in the context of belief bias in syllogistic reasoning, where such prior knowledge interferes with judgments of logical validity (Evans et al., 1983). For instance, when asked to evaluate the logical validity of the following syllogism: "No addictive things are inexpensive. Some cigarettes are inexpensive. Therefore, some cigarettes are not addictive," participants tend to rate this valid argument as invalid because it contradicts their prior belief that all cigarettes are addictive. In the context of the "lawyer-engineer" problem, these beliefs take the form of stereotype-driven expectations, for example the belief that someone who enjoys mathematical puzzles is likely to be an engineer. Here, belief strength is best understood as the strength of the stereotype, that is, the associative strength between a descriptive trait and a social category.

Typically, the information triggering the heuristic is verbal, as it needs to activate prepotent, automatic responses such as stereotypes in the base-rate neglect task. Consequently, measuring belief strength is costly and requires repetitive individual ratings from human participants to be correctly estimated. This has constrained researchers' ability to accurately measure and manipulate belief strength in reasoning tasks. Addressing this critical methodological gap is the primary aim of the current paper.

At their core, all heuristics-and-biases problems, such as the "lawyer-engineer" problem above, involve a conflict between (a) information that cues an intuitive, heuristic response (e.g., stereotype information suggesting Jack is an engineer) and (b) information that should be

considered according to normative principles of logic and probability (e.g., base-rate information indicating Jack is far more likely to be a lawyer).

To date, researchers have primarily manipulated two aspects of such problems. First, researchers have varied the alignment or conflict between heuristic and logical information by reversing base-rate probabilities to create "no-conflict" scenarios (e.g., changing the example to 996 engineers and 4 lawyers; De Neys et al., 2011; De Neys & Glumicic, 2008; Stupple & Ball, 2008; Stupple et al., 2011). Second, they have systematically varied the strength of logical information itself, for instance, using extremely skewed base-rates (e.g., 995 lawyers/5 engineers) versus moderately skewed base-rates (e.g., 700 lawyers/300 engineers; Pennycook et al., 2015). To our knowledge, however, no comparable systematic manipulations or precise measurements of belief strength have yet been developed.

On a purely methodological level, precise measurement of belief strength would enable researchers to control for variability across items. Indeed, current research typically assumes uniformly strong belief strength across stimuli, but unaccounted variability could significantly impact item reliability. For example, consider the widely used rapid-response base-rate neglect items (Pennycook et al., 2014), which use single adjectives to cue stereotypes instead of a lengthy individuating description (e.g., "There are 995 secretaries and 5 drummers. Person ‘L’ is loud. Is person ‘L’ more likely to be a secretary or a drummer?"). Each item assumes a consistent, high stereotype-driven belief strength. However, it is questionable whether each adjective points equally strongly toward one group compared to the other, across the different base-rate items. While variations in belief strength across items might not be important for experiments with large item sets that examine robust effects, they could undermine the validity of studies relying on small item sets or exploring subtle interactions (e.g., Bago & De Neys, 2020). This issue is particularly

relevant in studies that divide items into subsets, such as training or debiasing studies using pre-post designs, where an imbalance in belief strength between subsets of items could bias the results if not properly counterbalanced.

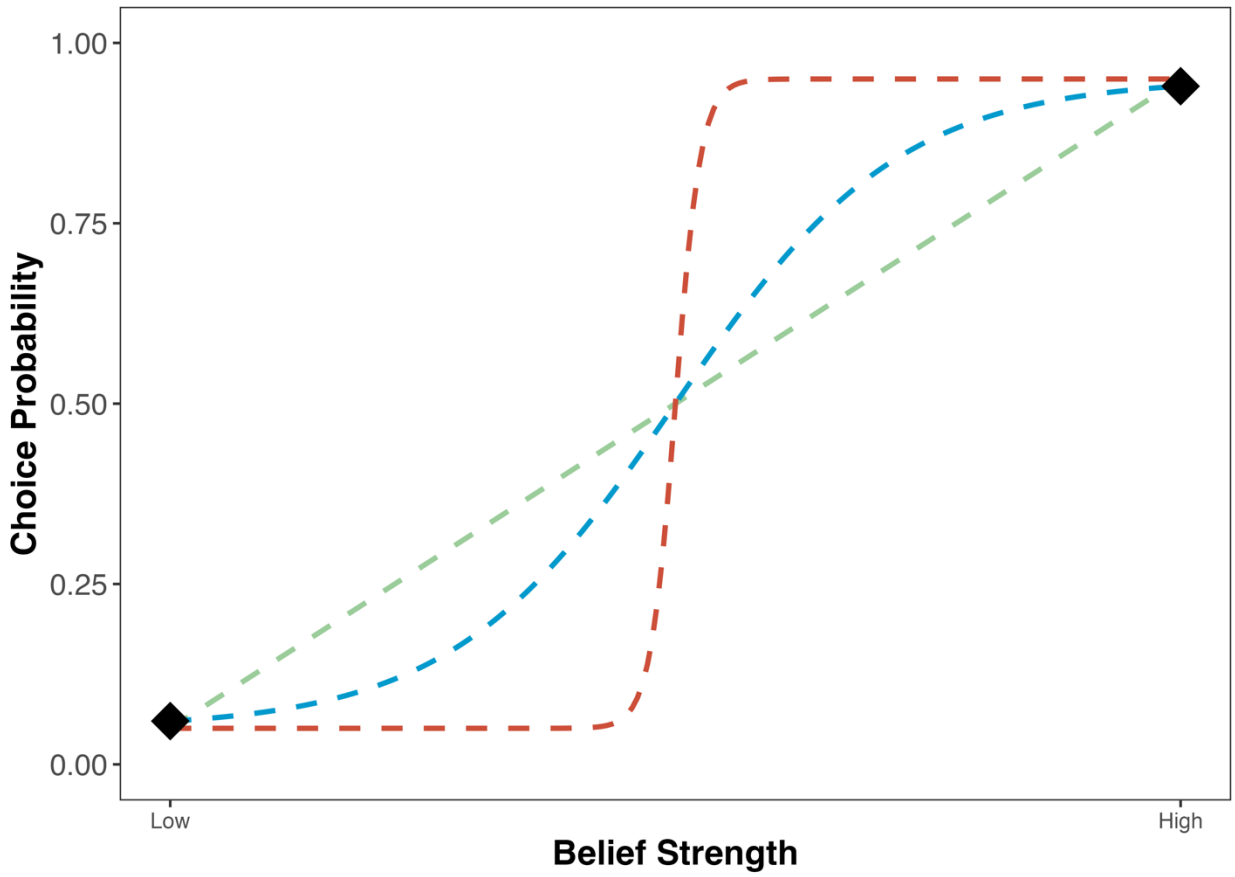

Figure 1. Illustration of different hypothetical response functions (linear, sigmoid, and step-like) linking belief strength to choice probability when only two belief strength levels (low and high, indicated by black diamonds) are used. This limited binary approach restricts researchers' ability to precisely characterize participants' underlying cognitive processes or strategies and differentiate among competing theoretical models.

On a more theoretical level, accurately measuring and manipulating belief strength is essential for understanding cognitive processes underlying heuristics-and-biases tasks. Currently, relying primarily on binary manipulations (e.g., conflict versus no-conflict) restricts researchers' ability to precisely examine how participants weigh heuristic versus logical information. Such limited stimulus variation also constrains computational modeling in reasoning research, as multiple competing models or functions could fit participants' data equally well, as illustrated in Figure 1. Parametric manipulations of heuristic information would thus offer greater sensitivity, enabling researchers to better distinguish between competing reasoning models (De Neys, 2023). This would align reasoning research with practices in other cognitive science fields—such as perception or reinforcement learning—where researchers routinely use precise stimulus manipulations to dissect underlying cognitive or computational processes.

To address these methodological and theoretical limitations, recent advances in natural language processing (NLP)—particularly large language models (LLMs)—offer a promising new approach. LLMs have successfully enabled the automation of verbal rating tasks with very high reliability (e.g., DiStefano et al., 2024; Le Mens et al., 2023; Ornstein et al., 2023). In this paper, we introduce a novel automated method using LLMs to systematically quantify belief strength in base-rate neglect items. We implement this approach using two high-performance LLMs: GPT-4 (OpenAI, 2023) and LlaMa 3.3-70B-Instruct (Grattafiori et al., 2024).

We validate this automated measure against human judgments by examining both explicit participant ratings and, importantly, actual choice patterns using these items. We further demonstrate the utility of our approach by applying it to widely used base-rate neglect stimuli from previous research (Pennycook et al., 2015), revealing substantial, previously unnoticed variability in belief strength. Finally, we provide an extensive, open-access database of over 100,000 unique

base-rate neglect items, following the rapid-response format from Pennycook et al. (2014). This comprehensive resource facilitates more precise, rigorous, and replicable future research. To support such use, we also provide the R package *baserater* (Beucler, 2025), which allows researchers to access the database and apply or adapt the method to their own experimental needs easily.

Experiment 1: Typicality Ratings Validation

Typicality ratings reflect the extent to which specific traits are perceived as representative of specific groups (e.g., "Nurses are typically kind"). In the present framework, typicality operationalizes belief strength by capturing how strongly a descriptive trait is associated with a given group. Experiment 1 aimed to validate our automated method for estimating typicality ratings using LLMs. We generated these ratings systematically with LLMs and compared them to human-generated ratings.

Methods

Base-Rate Task

We focus on the rapid-response base-rate task (Pennycook et al., 2014), widely used in reasoning research for its standardized and concise format. In each trial, a short vignette presents the composition of the sample (e.g., "This study contains nurses and politicians"), a description of a person with a neutral name and an adjective cueing a stereotype associated with one of the groups in the sample (e.g., "Person ‘L’ is kind"), and base-rate information (e.g., "There are 5 nurses and 995 politicians"). The task is to indicate which group the person most likely belongs to (e.g., "Is Person ‘L’ more likely to be a nurse or a politician?").

Estimation of Typicality Ratings Using LLMs

Stereotypes as Likelihood Estimates. In our base-rate neglect example, the information triggering the heuristic response—which we aim to quantify—is the stereotype embedded in the person's description. In this context, we treat stereotype strength as a specific type of belief strength—namely, a belief about category membership informed by stereotypical cues. Formally, a stereotype (e.g., "Nurses are kind") can be expressed as a conditional probability, or likelihood, of observing a specific trait (e.g., kind) given that one belongs to a specific group (e.g., nurse): p(trait|group). For instance, in a base-rate item such as: "There are 995 politicians and 5 nurses. Person 'L' is kind. Is person 'L' more likely to be a politician or a nurse?", both (a) the likelihood and (b) the base-rate information must be integrated to accurately estimate the probability. According to Bayes' theorem, the posterior probability that person 'L' is a nurse given that she is kind is expressed as follows:

$$p(H|D) = \frac{p(H)p(D|H)}{p(H)p(D|H)+p(\neg H)p(D|\neg H)} \quad (1)$$

where p(H) and p(¬H) are the base-rate probabilities that person 'L' is a nurse or a politician, respectively, and p(D|H) and p(D|¬H) are the likelihoods of observing the trait "kind" given that person 'L' is a nurse and a politician, respectively. To quantify the strength of the stereotype in each base-rate neglect item, we thus need to quantify this likelihood information for each group-adjective pair (e.g., for *nurse-kind* and *politician-kind*). In this context, an item with a high stereotype-driven belief strength will have a large disparity in likelihood between groups — such as an item where a trait is much more strongly associated with one category than the other.

In natural language, this likelihood is often captured by statements of "typicality." For instance, saying "Nurses are typically kind" expresses that the probability a person is kind, given

that they are a nurse, is high. Because LLMs are trained on extensive human-generated datasets, they inherently capture broad societal biases and perceptions, making them well-suited for estimating stereotype-based likelihoods. Prior research by Le Mens et al. (2023) demonstrated that LLMs can generate "typicality" ratings—quantifying how representative certain attributes are within specific categories—achieving near-perfect correlation ($r > 0.9$) with aggregated human judgments in domains such as literary genres and political affiliations.

Prompt Design. Building on this approach, we use LLMs to estimate the typicality of a given trait for a specific group. For each group-adjective pair (e.g., drummer–loud), we ask the LLM to rate how well the adjective (ADJECTIVE) describes the prototypical member of a group (GROUP). This approach directly follows Pennycook et al. (2015), who created base-rate items by asking participants to select traits they felt best represented prototypical group members. To help the LLM quantify this relationship clearly, we provided additional context through three illustrative examples (i.e., few-shot learning). Each LLM prompt consisted of two distinct sections: a context-setting instruction (system prompt) followed by detailed instructions and examples (user prompt). An example prompt is provided below (words in capital letters represent variables that change with each adjective or group):

System prompt

> "You are expert at accurately reproducing the stereotypical associations humans make, in order to annotate data for experiments. Your focus is to capture common societal perceptions and stereotypes, rather than factual attributes of the groups, even when they are negative or unfounded."

User prompt

> "Rate how well the adjective "ADJECTIVE reflects the prototypical member of the group "GROUP" on a scale from 0 ("Not at all") to 100 ("Extremely").

To clarify, consider the following examples:

1. ‘Rate how well the adjective "FUNNY" reflects the prototypical member of the group "CLOWN" on a scale from 0 (Not at all) to 100 (Extremely).’ In this example, you would likely give a high rating because the adjective ‘FUNNY’ closely aligns with the typical characteristics of a ‘CLOWN.’

2. ‘Rate how well the adjective "FEARFUL" reflects the prototypical member of the group "FIREFIGHTER" on a scale from 0 (Not at all) to 100 (Extremely).’ In this example, you would likely give a low rating because the adjective ‘FEARFUL’ diverges significantly from the typical characteristics of a ‘FIREFIGHTER.’

3. ‘Rate how well the adjective "PATIENT" reflects the prototypical member of the group "ENGINEER" on a scale from 0 (Not at all) to 100 (Extremely).’ In this example, you would likely give a moderate rating falling around the middle of the scale, because the adjective "PATIENT" neither closely aligns nor diverges significantly from the typical characteristics of an "ENGINEER."

Your response should be a single score between 0 and 100, with no additional text, letters, or symbols included."

LLMs Specification and Parameters. We used two high-performance LLMs: GPT-4 (version gpt-4-0613, queried in late April 2024 via OpenAI’s API) and the LlaMa 3.3-70B-Instruct model (accessed via Hugging Face’s hosted inference API). GPT-4 was chosen for its state-of-the-art performance, while LlaMa 3.3, a leading open-weight and publicly accessible model, was included to enhance transparency and reproducibility.

To account for variability in the model’s responses, we follow the general approach described by Le Mens et al. (2023) and generate 20 independent typicality scores for each group-adjective combination. To control variability in the model’s responses, we set two sampling parameters to their default values of 1: temperature, which controls randomness in token selection (higher temperature increases randomness, lower temperature produces more deterministic responses), and "Top-P", which means the model considers the entire posterior probability distribution of candidate tokens (lower values would limit sampling to the most probable tokens, whereas a value of 1 includes all tokens). In rare cases when the model fails to return a numerical response, we resubmit the prompt until obtaining a total of 20 valid ratings. The final likelihood

estimate p(adjective|group) is computed as the average of these 20 typicality scores divided by 100 to yield a probability ranging between 0 and 1. Note that individual scores were not recorded following aggregation since they showed very little variability. Supplementary Material 1 includes density plots and summary statistics describing the distribution of typicality ratings produced by each model across all group–adjective combinations.

Typicality Matrix. We selected our list of groups and adjectives from the base-rate items in Pennycook et al. (2015). The groups consisted of various professions chosen associated with common stereotypes (e.g., surgeon, artist, clown), while the adjectives reflected personality traits perceived as stereotypical (e.g., nerdy, arrogant, kind). Since our goal was to combine every group pairwise, we removed generic groups (poor people, rich people, girls, men, women) that could lead to ambiguous base-rate items creating class inclusion issues (e.g., "Is it more likely that the person is a man or a doctor?"). Additionally, we expanded the existing material by adding 14 additional groups (e.g., psychologist, soldier, fashion designer) and 41 new adjectives (e.g., naive, altruistic, shy), resulting in a total of 58 groups and 66 adjectives (see Supplementary Material 2 for the full list).

Typicality Ratings Validation Experiment

In Experiment 1, we collect individual human typicality ratings using the same procedure that we use with the LLMs (e.g., "Rate how well the adjective ‘nerdy’ reflects the prototypical member of the group ‘computer scientist’"). Our goal was to examine the correlation between human-generated ratings and those generated by the LLM.

Participants. Participants were recruited through the Prolific platform (www.prolific.com) and compensated £1.00 for 10 minutes of participation. Since stereotypes are likely culture-

specific, we recruited only native English-speaking North-American participants from the U.S. or Canada, consistent with the original study by Pennycook et al. (2015), which tested Canadian participants. A total of 50 participants were recruited (26 females; *M* age = 37.4, *SD* = 12.3), of whom 32% reported high school, 44% a bachelor's degree, 22% a master's degree, and 2% a PhD as their highest level of education.

Procedure. At the start of the experiment, participants reviewed the three examples provided to the LLM (see prompt above). Participants then completed a rating task for a subset of 100 group-adjective combinations sampled from the full LLM-generated typicality matrix. The subset was chosen to uniformly represent the entire range of typicality scores present in the full database (Experiment 1 GPT-4 typicality range: 11.3–95.8; full LLM database GPT-4 typicality range: 7.4–99.8). For each combination, participants rated how well an adjective described the prototypical member of a specified group (e.g., "Rate how well the adjective 'IDEALISTIC' reflects the prototypical member of the group 'WRITER'"). These combinations were presented in a random order. Participants provided ratings using a visual analog scale ranging from 0 ("Not at all") to 100 ("Extremely"), with the selected rating displayed numerically above the scale.

Results

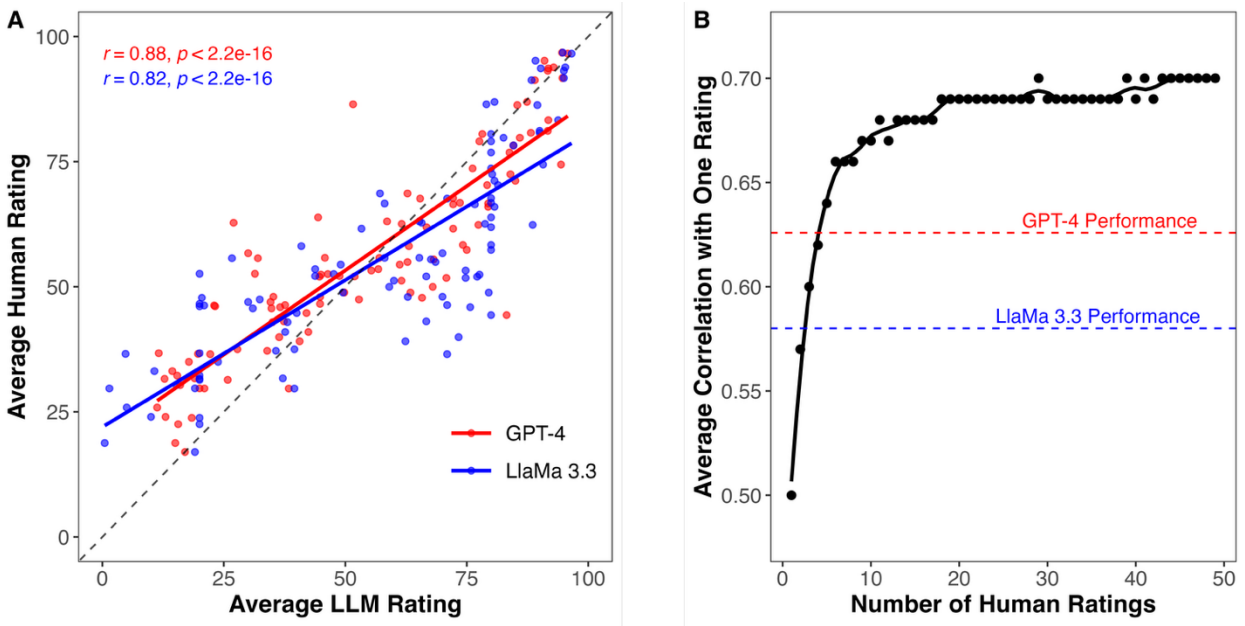

Figure 2. LLMs ratings are highly predictive of human typicality ratings. **a)** Relationship between the average LLMs rating and the average human typicality rating in Experiment 1. The solid lines represent linear model fits. Each point represents an adjective-group association. **b)** Average correlation between a single typicality rating and an aggregated measure of N human typicality ratings as a function of N. The dotted horizontal lines represent the average correlation achieved by each LLM. GPT-4 matches the performance of four to five human raters, while LlaMa 3.3 equals that of two to three human raters.

As shown in Figure 2A, typicality ratings from GPT-4 closely matched participants' average ratings in Experiment 1 ($r = 0.88$, $p < .001$), with LlaMa 3.3 showing slightly lower but still strong correlation ($r = 0.82$, $p < .001$). Note that averaging the typicality ratings produced by GPT-4 and LlaMa 3.3 produced only a very modest improvement (r = .89), indicating that combining both

models offers limited additional predictive accuracy over GPT-4 alone, which may be explained by the high correlation between the two LLMs' ratings (see Supplementary Material 3). LLM ratings can thus closely approximate the average human typicality rating in the context of stereotypes.

To better compare LLM performance to human raters, we use the Equivalent Number of Observations (ENO) score, following Le Mens et al. (2023). The ENO score quantifies how many human ratings are required to achieve predictive accuracy equivalent to that of a given model. Predictive accuracy, in this context, is defined as the average Pearson correlation between participants' typicality ratings and the ratings generated by the model.

To compute the ENO score, we performed a bootstrap analysis. Specifically, we correlated the ratings from one random participant (the holdout) with the average ratings from varying numbers of other randomly selected participants. We repeated this process 1,000 times for each sample size (ranging from 1 to 49 participants in our case). By comparing the average correlations from these bootstrap samples to the model's predictive accuracy, we determined the ENO score—the number of human ratings required to match the model's performance. A higher ENO score indicates greater predictive accuracy, as it corresponds to less noise and more consistent judgments, similar to how increasing sample size reduces variance and improves reliability.

The results are shown in Figure 2B. GPT-4 outperformed LlaMa 3.3, with an average correlation of $r = 0.63$ with a single human rating—equivalent to the performance of four to five aggregated human raters. In comparison, LlaMa 3.3 reached an average correlation of $r = 0.58$, matching the performance of two to three aggregated human raters. Importantly, increasing the

number of human raters only yields marginal gains, as it eventually plateaus at approximately r = 0.70.

Finally, to assess the robustness of our approach, we explored alternative methods for eliciting typicality ratings from the LLMs by varying model settings and using alternative prompts. Overall, the correlation with human typicality ratings was minimally affected by changes in settings or prompts (LlaMa 3.3: $r = 0.78–0.82$; GPT-4: $r = 0.85–0.89$). The detailed results of these analyses are reported in Supplementary Material 4.

Experiment 2: Measuring Stereotype Strength in Base-Rate Neglect Items

In Experiment 1, we validated our approach by demonstrating that LLM-generated typicality ratings closely align with explicit human judgments. Although this correlation provides strong evidence for the validity of our automated method, a critical next step is to examine whether these ratings can predict actual behavioral choices. To address this, we compute a measure of stereotype strength—based on the relative typicality between two groups—and examine its predictive validity in a base-rate neglect task using our newly created item database in Experiment 2.

Methods

Stereotype Strength Measure

A typical rapid-response base-rate neglect item from Pennycook et al. (2014) includes two groups (e.g., nurses and politicians), one descriptive adjective (e.g., "kind"), and two corresponding typicality ratings—one rating for how typical the adjective is of each group. Because typicality ratings were provided on a 0–100 scale, we converted them to likelihoods by dividing each rating by 100. Likelihoods of zero were offset (1e-6) to allow calculations of the log

ratio. This included 0.55% of LlaMa 3.3 ratings and 0% of GPT-4 ratings. To ensure this post-processing did not affect our conclusions, we examined the correlations between human and LlaMa ratings before and after excluding raw zero ratings; this exclusion did not impact the correlation (see Supplementary Material 5). To quantify how strongly a given adjective favors one group over the other, we computed stereotype strength as the logarithm of the ratio between these two group-adjective likelihoods: $log\ \frac{p(A|G_1)}{p(A|G_2)}$ (2)

where p(A|G) is the likelihood derived from the typicality rating for adjective A given group G.

This log ratio provides a symmetrical measure that quantifies the strength of association between the trait and each group. A log ratio of zero indicates equal likelihood of the trait in both groups, meaning the adjective is uninformative with respect to the two groups. Positive values suggest a stronger association with the first group, while negative values indicate a stronger association with the second group.

For instance, consider the two following items from Pennycook et al. (2015): *computer programmers–construction workers–nerdy* and *consultants–aerobics instructors–helpful*. We might imagine that the likelihood of being perceived as nerdy, given that one is a computer programmer, is high, p(nerdy|computer programmer) = 0.8, while it is much lower for a construction worker, p(nerdy|construction worker) = 0.1. This yields a log ratio of log(0.8/0.1) = 2.08, which strongly favors the computer programmers group over the construction workers group. Conversely, we might imagine that both p(helpful|teacher) and p(helpful|doctor) are high, say 0.8, which results in a log ratio of log(0.8/0.8) = 0. This means that, in this case, the item will fail to elicit a strong heuristic response.

Construction of Base-Rate Item Database

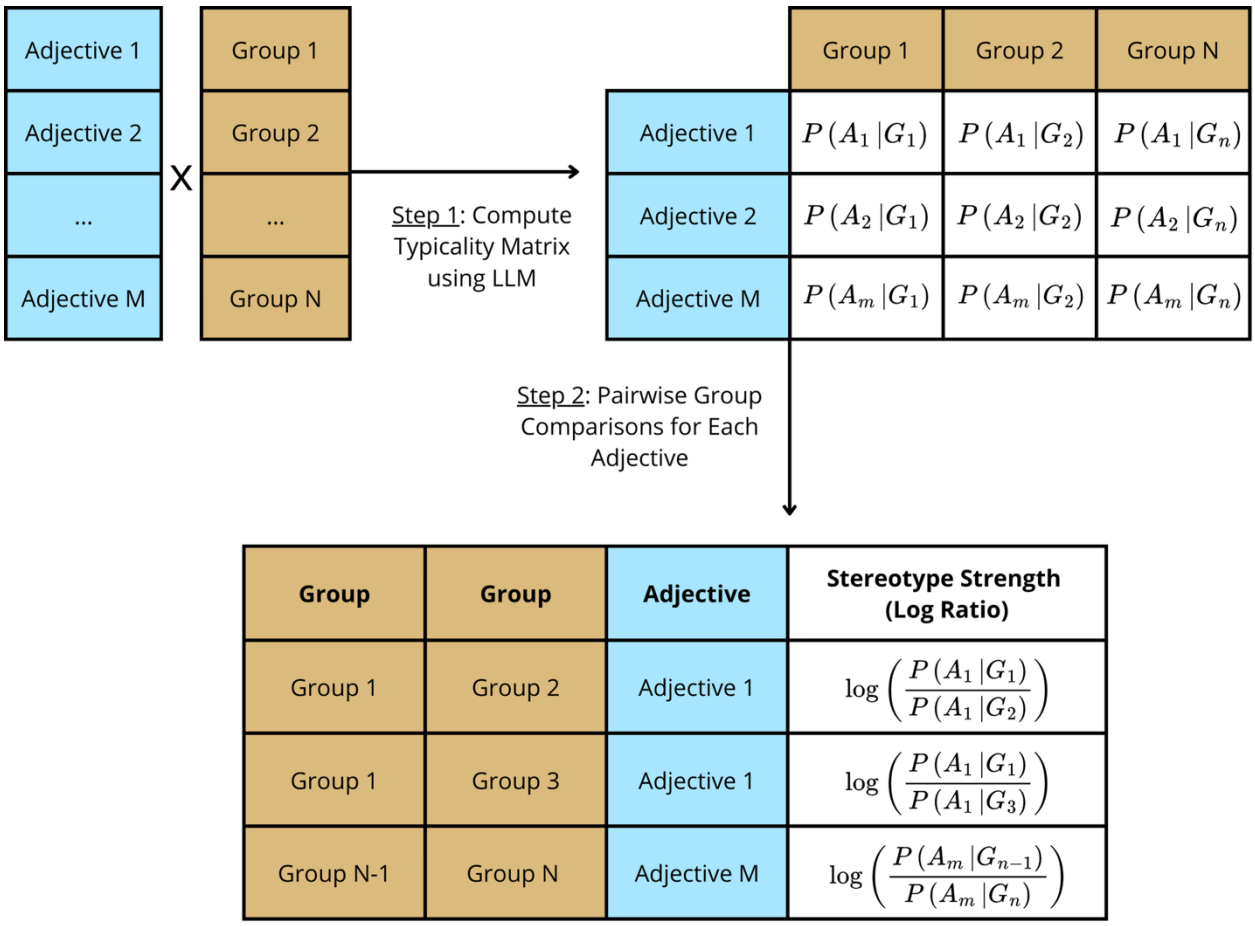

Figure 3. Overview of the base-rate item database pipeline. The typicality of each adjective is first computed for each group using an LLM (step 1) to create the typicality matrix. Each typicality score can be interpreted as the conditional probability of having a specific trait given that one belongs to a particular group (e.g., the probability of being kind given that one is a computer scientist). Next, for each pairwise combination of groups with one adjective (e.g., clown-nurse-funny), the stereotype strength is determined by calculating the log ratio of the two typicality scores/conditional probabilities, to create the final base-rate item database (step 2).

To create our base-rate item database, we first computed the typicality rating of each adjective for each group using the LLM method described above, separately for GPT-4 and LlaMa 3.3. This resulted in a 66-adjective × 58-group typicality matrix. We then created individual base-rate items, each consisting of two groups and one adjective (e.g., *clown-nurse-funny*). We generated every possible pairwise combination of groups (1,653 combinations) and computed the stereotype strength for each of the 66 adjectives, yielding a total of 109,098 base-rate items. The full pipeline is described in Figure 3.

Base-Rate Neglect Validation Experiment

In Experiment 2, we use our newly created item database in the rapid-response base-rate neglect task (Pennycook et al., 2014), where we systematically vary stereotype strength to see whether it accurately predicts participants' choices. Note that we rely on GPT-4 measures, as they proved to be more correlated with human judgments than those derived from LlaMa 3.3 (see Experiment 1 results).

Participants. We recruited 151 native English-speaking participants from the U.S. or Canada (74 females; M age = 39, SD = 11.9) via Prolific, who were compensated £3.00 for 30 minutes. 33% reported high school, 48% a bachelor's degree, 15% a master's degree, and 4% a PhD as their highest education level.

Procedure. Participants solved 240 base-rate neglect problems, divided into 4 blocks of 60 trials. Each trial began with a fixation cross presented during 500 ms, followed by the sample composition (e.g., "This study contains computer programmers and hippies."), a brief description of an individual with a neutral name and adjective (e.g., "Person 'G' is nerdy."), and the base-rate information (e.g., "There are 50 computer programmers and 950 hippies."). Participants indicated

the most likely group membership by pressing ‘C’ or ‘N’ to select the left or right response, respectively. Confidence ratings were collected after each response.

Items were dynamically and randomly sampled for each participant from the base-rate neglect database to ensure uniform coverage of stereotype strength. To achieve this, we divided the full set of items into predefined bins based on the difference between the two typicality scores (e.g., 0–14, 14–28, …, up to 84—the highest observed difference in the dataset). These bins allowed us to sample items across the full range of stereotype strength values in a controlled manner. Participants thus saw different, unique base-rate items, matched in stereotype strength according to these stereotype strength difference bins—a procedure made possible by the large number of items available in our database. In parallel, we used ten different base-rate ratios ranging from 50/950 to 950/50 in steps of 100, so that some problems involved balanced base rates while others were more extreme. Each base-rate ratio was paired with an equal number of items from each stereotype-strength bin to maintain balanced coverage across the full range of both dimensions.

While item selection was based on typicality score differences, our analyses rely on the absolute log ratio of category likelihoods, which provides a more theoretically grounded measure of stereotype strength (see above). Since our primary interest is the relationship between stereotype strength and participant choices, we collapsed our analyses across base-rate values. Note that when stereotype strength is high but base-rates are weak for a given response option, choosing the stereotype-consistent option is not always the normatively correct response under Bayes’ rule. However, our focus here is not on normative accuracy, but on how strongly stereotype strength predicts human choices.

Results

LLMs' typicality ratings closely match human judgments, but a more critical test is whether our stereotype-driven belief strength measure (i.e., the log ratio of typicality ratings) predicts actual human decisions in a base-rate neglect experiment (Experiment 2). In all analyses, response choice was coded as 1 when participants selected the option on the left and 0 otherwise. Stereotype strength was computed as the log ratio of typicality values in favor of the left option: $\log p(\text{adjective} \mid \text{group left}) / p(\text{adjective} \mid \text{group right})$. Figure 4 illustrates the proportion of choices as a function of stereotype strength, aggregated across all base-rate conditions. For visualization purposes, we binned the signed log ratio—calculated in favor of the response option presented on the left—into 10 intervals of width 0.5, ranging from –2.5 to 2.5. The figure shows that participants' choices are well predicted by our stereotype strength measure.

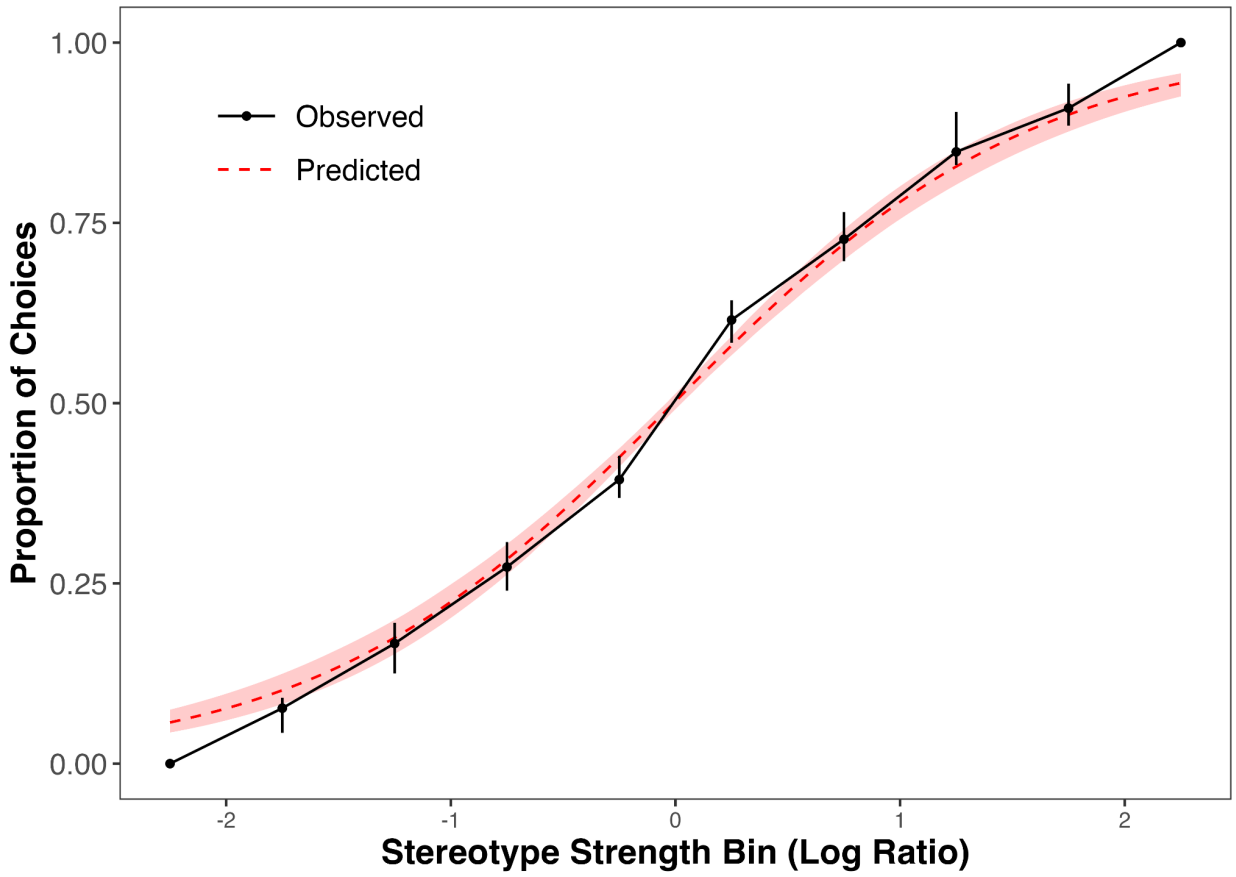

Figure 4. Proportion of choices as a function of the GPT-4 stereotype-driven belief strength measure (binned). Black dots represent the observed median proportion of choices for each log-odds bin of stereotype strength, with error bars indicating bootstrapped confidence intervals. The dashed red line represents predictions from a mixed-effects logistic regression model, and the shaded red ribbon shows the bootstrapped confidence interval for these predictions.

To test this statistically, we built a generalized mixed-effects logistic model, using response choice (left = 1, right = 0) as the dependent variable and stereotype strength—measured as log(p(adjective|group left) / p(adjective|group right))—as a fixed effect. We included uncorrelated random intercepts and slopes for participants in the model. Model comparisons based on the

Bayesian Information Criterion (BIC) indicated that this random-effects structure provided the best fit to the data. As shown in Figure 4, results indicate that stereotype strength is a statistically significant and strong predictor of participants' choices, regardless of base-rate values, $OR = 3.49$, $p < .001$, 95% CI = [3.06, 3.97].

Analysis of Existing Base-Rate Items

The previous analysis showed that our stereotype-based belief strength measure effectively predicts participants' choices in a controlled base-rate neglect task, highlighting its validity. However, an important question is how strongly the items from prior research—designed to evoke stereotypical responses—are characterized by our metric. Do they score uniformly high on our stereotype strength measure?

In this section, we leverage our method to quantify the stereotype strength of our selection of base-rate neglect items previously used by Pennycook et al. (2015). Since we excluded items involving overly generic groups (e.g., "poor people" or "girls"; see Method section above), our analysis focuses on 88 of the original 132 items. We then compare these to the items in our newly created base-rate item database.

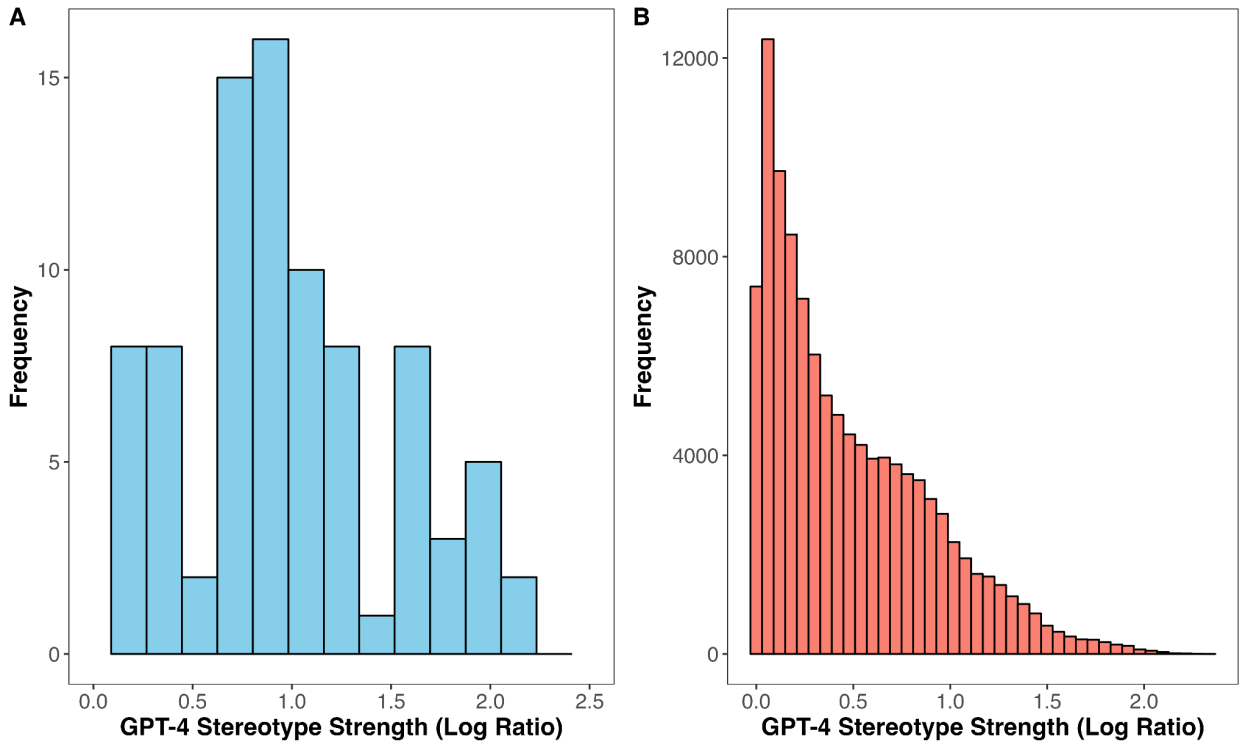

Figure 5. Stereotype strength distributions based on the GPT-4 ratings. **a)** Stereotype strength distribution in existing base-rate items from Pennycook et al. (2015) present in our database (n = 88 items). **b)** Stereotype strength distribution in the new base-rate item database (n = 109,098 items). Note that the log ratios in the database are always positive by construction, since we ensured that the item with the highest typicality ratings was always in the numerator of the ratio.

As shown in Figure 5A, none of the existing base-rate items elicit a stereotype in the opposite direction. In other words, no item has a negative log ratio value, where the group favored by the adjective differs from the one assumed by the item. Nonetheless, the stereotype strength of these items varies widely, spanning from 0.02 to 2.17 in absolute log ratio values ($M = 0.97$, $SD = 0.54$). Supplementary Material 6 shows the stereotype strength distributions based on ratings from LlaMa 3.3.

To illustrate the potential behavioral implications of this variability, we used our generalized logistic mixed-effects model fitted on the data from Experiment 2 (i.e., the base-rate neglect experiment). The model allows us to predict response choice as a function of the GPT-4 stereotype strength in the existing items with a very good fit (see Figure 4). For the weakest item in the existing dataset (*consultant – aerobics instructor – helpful*), the log ratio is 0.02, predicting a 51% probability of selecting "consultant". This suggests that participants would be at chance level when choosing based on the stereotype (corresponding to the midpoint of the sigmoid curve in Figure 4). Conversely, for the strongest item (*high school coach – librarian – loud*), the log ratio is 2.17, predicting that "high school coach" would be selected 94% of the time. Overall, for the average log ratio value in existing items (0.97), the model predicts that the participant would choose this group 77% of the time. This wide variation in stereotype strength among base-rate items could thus be problematic for experimental validity, as weaker stereotypes may inconsistently trigger heuristic responses, undermining reliability and replicability of results.

Overview of the Base-Rate Item Database

Figure 5B shows the full distribution of stereotype strengths within the database, while Supplementary Material 7 summarizes key statistics separately for the LlaMa 3.3 and GPT-4 ratings. As expected, most adjectives do not favor either of the two groups, resulting in a stereotype strength distribution heavily skewed to the right. However, given the substantial size of the database (109,098 items), it still contains many high-strength stereotype items. For instance, while Figure 5B may give the impression that very few items exceed a (high) GPT-4 stereotype strength of 1.5, the database actually contains 2,446 such items, providing a rich dataset of high stereotype items. For illustration, Figure 6 shows an example of all 1,653 possible items that can be created

based on one of our 66 adjectives ("arrogant"). In addition, Supplementary Material 8 provides examples of these items along with their associated typicality ratings and stereotype strengths from both GPT-4 and LlaMa 3.3.

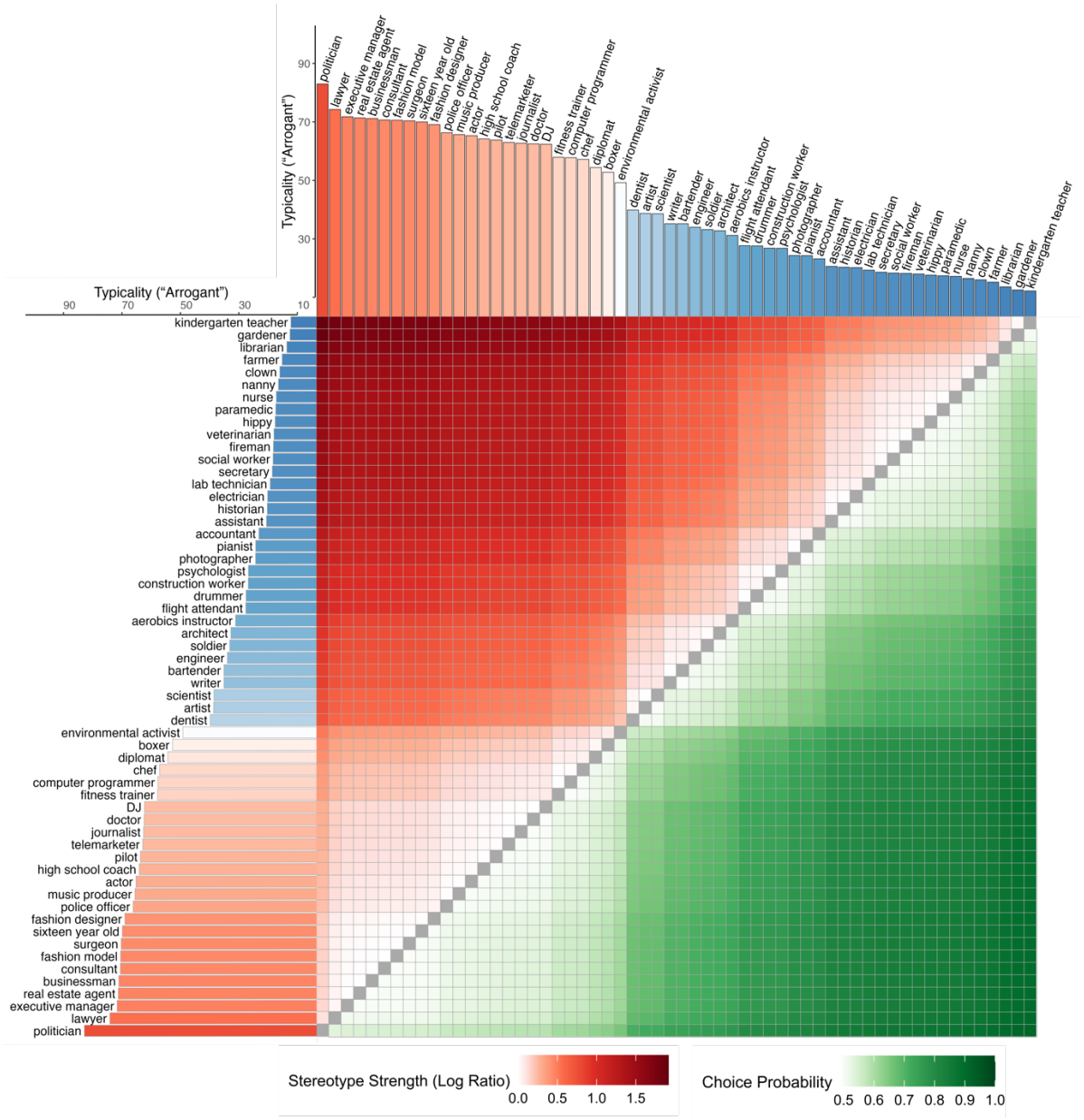

Figure 6. Illustration of all possible items for the adjective "Arrogant." The upper part of the matrix shows stereotype strength, computed as the log-ratio of GPT-4 typicality ratings (column / row). Higher values (darker red) indicate that the column group is rated as more "arrogant" than the row

group (e.g., "politician" over "kindergarten teacher" in the top-left cell). The lower part of the matrix shows predicted human choice probabilities from the mixed-effects logistic model fit to the base-rate neglect task (Experiment 2), collapsed across base-rate values. Higher values (darker green) reflect a greater likelihood of choosing the row group over the column group based on the stereotype (e.g., "politician" over "kindergarten teacher" in the bottom-right cell). The bar plots on the margins show the raw GPT-4 typicality ratings across each group, where taller and redder bars indicate higher typicality scores for that group.

Discussion

In this paper, we introduce an automated method to measure and quantify belief strength in heuristics-and-biases research using LLMs focusing on the popular base-rate neglect task. Using this approach, we created a comprehensive, open-access database containing over 100,000 unique base-rate neglect items. Importantly, the database spans a wide range of stereotype-driven belief strengths, allowing researchers to systematically vary the strength of the corresponding heuristic response in a precise manner.

We validated this method by comparing it with human judgments, using both participant typicality ratings (i.e., how typical a description is of a specific group; Experiment 1) and their actual choices on the newly generated base-rate items (Experiment 2). Our results show that our automated measure of belief strength, generated with LLMs, is highly correlated with human typicality judgments. More importantly, the LLM-derived belief strength measure strongly predicts human choices in a base-rate neglect task, so that the stronger the belief strength induced by a stereotype, the more likely participants are to choose the option that aligns with the stereotype.

Overall, GPT-4 outperformed the publicly available LlaMa 3.3 model, as GPT-4's ratings showed higher correlations with human typicality judgments. Additionally, averaging ratings from both models did not yield substantial improvements in performance. While we report results from LlaMa 3.3 to support research transparency and reproducibility, we recommend using GPT-4 ratings for applications where performance is critical.

Our LLM-derived measure is thus a highly cost-effective proxy for the average belief strength across a population. Measuring belief strength using human raters involves averaging judgments from multiple participants. However, the question of how many ratings are needed to produce accurate estimates of belief strength is unclear. Comparing our LMM and human rater approaches, we found that increasing the number of raters beyond a certain threshold yields diminishing returns (see Figure 2B). Indeed, we observed that interindividual differences in stereotype perception limited the precision achievable by any aggregate measure, human or automated. A more precise measure would involve obtaining individual ratings directly from each participant performing the task, with the need to add an additional session following the main experiment dedicated to estimating those associations.

However, human-based methods are often impractical due to substantial constraints related to budget, time, and resources, especially when evaluating large numbers of stimuli. Although imperfect, our automated approach effectively addresses these practical challenges by striking an efficient balance between precision and scalability, providing reliable estimates of average belief strength at the group level. Consequently, this approach allows researchers to quickly generate extensive datasets with standardized measures, which would otherwise be difficult to conduct using traditional human-based methods alone. In our case, we generated 3,828 typicality ratings to build the 109,098-item database. Each rating required a single 418-token API call (398 input

tokens and 20 output tokens to generate and average 20 individual LLM ratings). The total cost for this process was approximately $54.00 using GPT-4 and $1.42 using LlaMa 3.3. Using the more recent OpenAI GPT-4o model would have cost only about $4.60 for the same task. By comparison, collecting equivalent human data on Prolific, which required five individual ratings per item to achieve similar performance to GPT-4, would have cost about $320 in total, assuming a pay rate of $7.72 per hour per participant and a 30% platform fee. Overall, the LLM-based approach is roughly six times less costly with GPT-4, 70 times less costly with GPT-4o, and 225 times less costly with LlaMa 3.3 than human ratings.

Finally, we applied this approach to widely used base-rate neglect items from Pennycook et al. (2015), which were all designed to elicit a strong heuristic response by presenting descriptions that strongly favored one group over the other. Our analysis revealed substantial variation in belief strength driven by stereotype content across these items. Specifically, some items exhibited high stereotype-based belief strength, while others were comparatively less effective in eliciting a strong belief. This variability has important consequences for reasoning research. In the context of a base-rate neglect experiment, for instance, this suggests that existing items may not equally trigger heuristic responses, potentially affecting the reliability and replicability of findings across studies.

By systematically quantifying belief strength, our approach provides a way to better control for this variability. This enables one to refine item selection or to dynamically adjust belief strength to match experimental needs. To facilitate its broader use, we also provide the R package *baserater* (Beucler, 2025), which allows researchers to access the database and evaluate new base-rate items (e.g., in a different language or with longer stereotypical descriptions). Users can choose a preferred large language model, adjust generation parameters, and customize the prompt to explore

whether different models, settings, and prompt formulation better predict the human typicality ratings from Experiment 1. The package can also be used to create a new base-rate item database from scratch, for instance in a different linguistic and cultural context. Access to the LLMs requires an API token from a supported inference provider, such as the open-source platform Hugging Face (https://huggingface.co). All functionalities are documented in the package, which includes a tutorial to guide users through typical use cases.

In this paper, we focused on short base-rate items that are widely used in reasoning research for precise measurement of reaction times or neuroimaging due to their standardized format. However, the same procedure can, in principle, be applied to base-rate paradigms involving richer, more complex descriptions. For example, it can be applied to items similar to the classic lawyer–engineer problem, where participants receive an extended personality vignette for each profession. Such scenarios preserve the core structure of base-rate reasoning while offering greater ecological validity, as individuals must evaluate multiple pieces of information rather than a single cue. This is also a particularly promising direction because such richly structured vignettes are notoriously labor-intensive to norm with human ratings, whereas automated LLM-based ratings provide a scalable alternative.

Similarly, our approach could be adapted to other heuristics-and-biases tasks, where a verbal description is assumed to activate a belief signal that biases participants' responses. One natural candidate is the conjunction fallacy demonstrated in the well-known "Linda problem" (Tversky & Kahneman, 1983). In this task, a stereotypical description triggers a heuristic response conflicting with a fundamental probability rule—that the probability of two events occurring together cannot exceed the probability of either event alone. In syllogistic reasoning tasks featuring a conflict

between logical validity and the believability of the conclusion, this approach could be used analogously to compute the belief strength of the conclusion independently of its logical status.

Conceptually, extending our method to more complex, multi-cue reasoning tasks requires specifying, for each paradigm, (a) which aspects of the description shape participants' belief-based response and (b) what linguistic unit the LLM should rate. We see two complementary strategies for specifying this rating unit. A first, holistic approach treats the entire vignette as the unit to be rated: the LLM is asked how well the full description fits each available response option, yielding typicality scores that can be turned into a belief-strength index via a log ratio of typicality ratings. A second, more granular approach decomposes the vignette into constituent traits (e.g., "enjoys mathematical puzzles", "shows no interest in political issues"), computes belief strength for each trait separately as a log ratio of typicality ratings, and then aggregates these component-wise measures to approximate the overall belief strength. For instance, one can sum them on the log-odds scale, under a standard conditional-independence assumption that traits provide independent evidence given the group.

In either case, both strategies require careful prompt design and, critically, new validation work to test whether the resulting belief-strength measures correlate with human judgments and predict behavior in the corresponding tasks. Future research should also compare these two approaches directly to determine which measure provides the most appropriate representation of belief strength in light of the specific research questions and stimuli used. More broadly, any paradigm in which a verbal description is assumed to activate a prepotent response can be treated within the same framework: identify the relevant belief signal, decide whether to model it at the holistic or individual component level, compute LLM ratings for the corresponding linguistic units, integrate these ratings into a quantitative belief-strength parameter, and validate that

parameter against human ratings and choice behavior. Our method thus provides a generalizable framework for investigating situations in which heuristic responses triggered by verbal descriptions conflict with formal logical principles.

Our research has some limitations. First, although we used a prompt directly inspired by previous research on LLMs' typicality ratings (Le Mens et al., 2023) as well as base-rate neglect item construction (Pennycook et al., 2015), a more data-driven approach could enhance the predictive power of our belief strength measure. Notably, automatic prompt-engineering techniques (e.g., Abraham et al., 2025), which use LLMs to iteratively craft high-quality prompts, could further improve the prediction accuracy of our typicality measure, given that prompt selection can significantly affect performance outcomes (Weber & Reichardt, 2023). However, note that variations in our prompt and LLM settings resulted in only small differences in performance (see Supplementary Material 4).

Our current base-rate database has only been validated with participants from the U.S. and Canada. Although some stereotypes—such as "clowns are funny"—may appear broadly generalizable, others could be more culture-specific and thus fail to trigger heuristic responses in different populations. It will therefore be important for future work to consider potential intercultural differences. To ensure that belief strength measures remain applicable across cultures, we recommend refining prompts to better align with the target cultural context (e.g., Kovač et al., 2023), using fine-tuned LLMs trained specifically on culturally distinct datasets (e.g., Chan et al., 2024), or employing culturally adapted LLMs trained on augmented survey data (Li et al., 2024).

By harnessing automated methods and large-scale data generation, our approach provides researchers more control and precision in quantifying belief strength. This methodological advancement not only addresses critical limitations in current research, such as unnoticed

variability in belief strength across items, but also substantially improves the ability to distinguish among competing cognitive models of heuristic reasoning. Our open-access database containing over 100,000 systematically rated items, validated against human judgments, represents a powerful resource that can significantly enhance the rigor, replicability, and theoretical clarity of future heuristics-and-biases research.

Supplementary Material

The online supplementary material is available via the Open Science Framework at: https://doi.org/10.17605/OSF.IO/JCEYD.

Declarations

Funding

This research was supported by the Agence Nationale de la Recherche (ANR; ANR-23-AERC-0006 to ZP, ANR-23-CE28-0004-01 to WDN) and the Economic and Social Research Council (UKRI, ESRC; ES/V00378X/1 to LC).

Conflicts of Interest/Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Ethics Approval

This study was performed in line with the principles of the Declaration of Helsinki. The study was approved by the CER U-Paris ethics committee (Université de Paris).

Consent to Participate

All participants provided informed consent before taking part in the study.

Consent for Publication

Not applicable.

Availability of Data, Materials and Code

All data, materials and analysis scripts are available at: https://doi.org/10.17605/OSF.IO/JCEYD. The *baserater* package can be downloaded from CRAN at: https://CRAN.R-project.org/package=baserater.

Acknowledgements

We thank Gordon Pennycook for sharing pretest data on typicality ratings with us, as well as Laura Charbit and Nicolas Beauvais for their comments on earlier versions of the manuscript.

Open Practices Statement

All data, materials, and analysis scripts are publicly available at: https://doi.org/10.17605/OSF.IO/JCEYD. The *baserater* package can be downloaded from CRAN at: https://CRAN.R-project.org/package=baserater. None of the reported studies were preregistered.

References

Abraham, L., Arnal, C., & Marie, A. (2025). *Prompt Selection Matters: Enhancing Text Annotations for Social Sciences with Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2407.10645

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Bergus, G. R., Chapman, G. B., Gjerde, C., & Elstein, A. S. (1995). Clinical reasoning about new symptoms despite preexisting disease: Sources of error and order effects. *Family Medicine*, *27*(5), 314–320.

Beucler, J. (2025). *baserater: An R package using large language models to estimate belief strength in reasoning.* R package version 0.1.2. https://doi.org/10.32614/CRAN.package.baserater

Beucler, J., Purcell, Z., Charles, L., & De Neys, W. (2025). *Data, materials, and analysis scripts for Using Large Language Models to Estimate Belief Strength in Reasoning*. [OSF repository]. Open Science Framework. https://doi.org/10.17605/OSF.IO/JCEYD

Chan, A. J., García, J. L. R., Silvestri, F., O'Donnell, C., & Palla, K. (2024). *Enhancing Content Moderation with Culturally-Aware Models*. arXiv. https://doi.org/10.48550/arXiv.2312.02401

De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, *46*, e111. https://doi.org/10.1017/S0140525X2200142X

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS One*, *6*(1), e15954. https://doi.org/10.1371/journal.pone.0015954

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

DiStefano, P. V., Patterson, J. D., & Beaty, R. E. (2024). Automatic Scoring of Metaphor Creativity with Large Language Models. *Creativity Research Journal*, 1–15. https://doi.org/10.1080/10400419.2024.2326343

Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*(3), 295-306. https://doi.org/10.3758/BF03196976

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The LlaMa 3 Herd of Models*. arXiv. https://doi.org/10.48550/arXiv.2407.21783

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251. https://doi.org/10.1037/h0034747

Kovač, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P. F., & Oudeyer, P.-Y. (2023). *Large Language Models as Superpositions of Cultural Perspectives*. arXiv. https://doi.org/10.48550/arXiv.2307.07870

Le Mens, G., Kovács, B., Hannan, M. T., & Pros, G. (2023). Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences*, *120*(49), e2309350120. https://doi.org/10.1073/pnas.2309350120

Li, C., Chen, M., Wang, J., Sitaram, S., & Xie, X. (2024). Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, *37*, 84799–84838. https://doi.org/10.48550/arXiv.2402.10946

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 Technical Report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Ornstein, J. T., Blasingame, E. N., & Truscott, J. S. (2025). How to train your stochastic parrot: Large language models for political texts. *Political Science Research and Methods*, *13*(2), 264-281. https://doi.org/10.1017/psrm.2024.64

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*(1), 1–10. https://doi.org/10.3758/s13421-013-0340-7

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., Newton, C., & Thompson, V. A. (2022). Base-rate neglect. In *Cognitive illusions* (pp. 44–60). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003154730-5/base-rate-neglect-gordon-pennycook-christie-newton-valerie-thompson

Stupple, E. J. N., & Ball, L. J. (2008). Belief–logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking & Reasoning*, *14*(2), 168–181. https://doi.org/10.1080/13546780701739782

Stupple, E. J., Ball, L. J., Evans, J. S. B., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. Journal of Cognitive Psychology, 23(8), 931-941. https://doi.org/10.1080/20445911.2011.589381

Thompson, W. C., & Schumann, E. L. (2017). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. In *Expert evidence and scientific proof in criminal trials* (pp. 371–391). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315094205-15/interpretation-statistical-evidence-criminal-trials-william-thompson-edward-schumann

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293. https://doi.org/10.1037/0033-295X.90.4.293

Weber, M., & Reichardt, M. (2023). *Evaluation is all you need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer using Open Models*. arXiv. https://doi.org/10.48550/arXiv.2401.00284

Supplementary Material

1. Typicality Ratings Across Group–Adjective Pairs

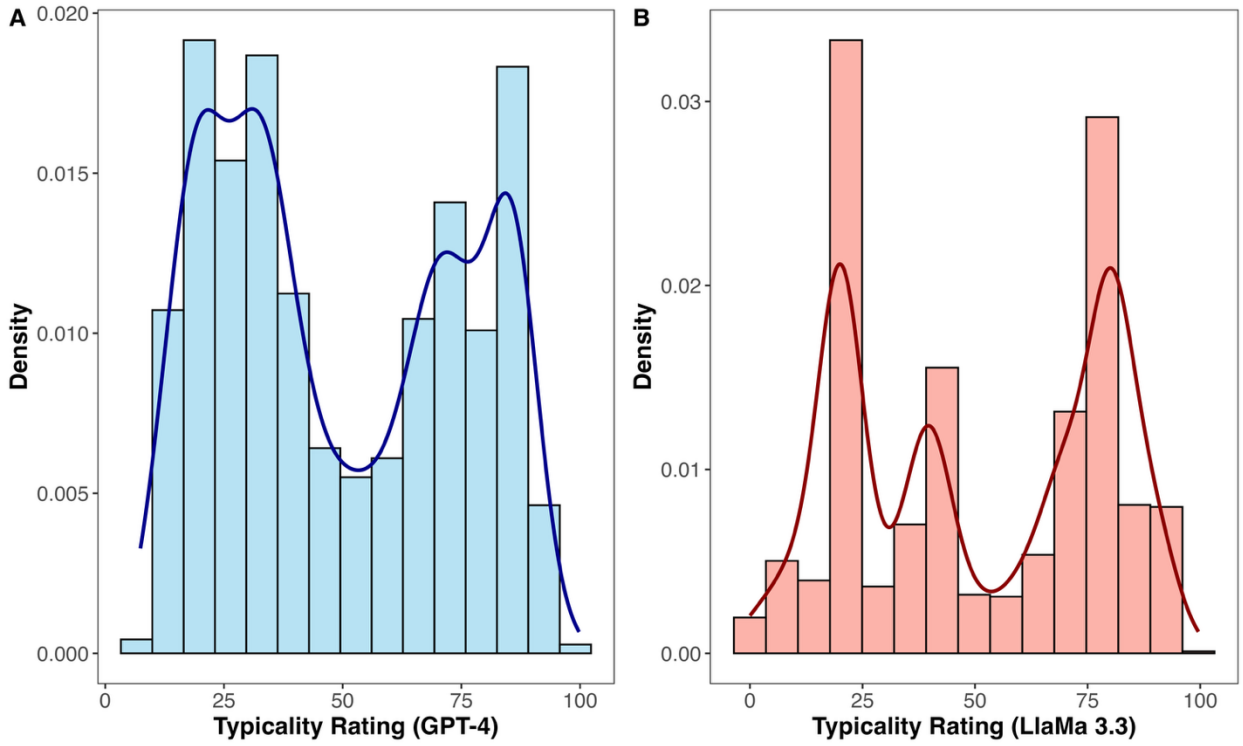

Figure S1. Distribution of typicality ratings across all group–adjective combinations for GPT-4 and LlaMa 3.3.

Table S1

Summary statistics of the typicality ratings across all groups–adjective combinations for GPT-4 and LlaMa 3.3.

Model	Mean	Median	SD	Min	Max
GPT-4	49.8	44.7	25.2	7.4	99.8
LlaMa 3.3	50.6	46.0	27.8	0.0	99.6

Note. SD = standard deviation; *Min* = minimum; *Max* = maximum.

2. Groups and Adjectives Used in the Database

Table S2

List of groups and adjectives used in the database.

Group	New Group	Adjective	New Adjective
farmer	Old	intelligent	Old
computer programmer	Old	arrogant	Old
flight attendant	Old	nerdy	Old
high school coach	Old	kind	Old
dentist	Old	loud	Old
lawyer	Old	careful	Old
engineer	Old	argumentative	Old
real estate agent	Old	persuasive	Old
accountant	Old	immature	Old
surgeon	Old	active	Old
architect	Old	funny	Old
librarian	Old	disorganized	Old
lab technician	Old	dishonest	Old
artist	Old	gentle	Old
consultant	Old	sensitive	Old
scientist	Old	creative	Old
nanny	Old	helpful	Old
boxer	Old	strong	Old
paramedic	Old	brave	Old
businessman	Old	bossy	Old
secretary	Old	unconventional	Old
executive manager	Old	quiet	Old
assistant	Old	organized	Old
nurse	Old	reliable	Old
writer	Old	ambitious	Old
telemarketer	Old	charming	New
clown	Old	confident	New
fireman	Old	efficient	New
pianist	Old	friendly	New
doctor	Old	generous	New
hippy	Old	naive	New
construction worker	Old	witty	New

Group	New Group	Adjective	New Adjective
gardener	Old	empathetic	New
aerobics instructor	Old	stubborn	New
sixteen year old	Old	trustworthy	New
politician	Old	meticulous	New
kindergarten teacher	Old	inventive	New
drummer	Old	charismatic	New
chef	Old	reserved	New
bartender	Old	altruistic	New
pilot	Old	original	New
social worker	Old	impulsive	New
veterinarian	Old	zealous	New
journalist	Old	rational	New
police officer	New	idealistic	New
electrician	New	conservative	New
fitness trainer	New	solitary	New
psychologist	New	passionate	New
actor	New	adventurous	New
historian	New	cautious	New
DJ	New	extravagant	New
diplomat	New	jovial	New
environmental activist	New	cooperative	New
music producer	New	attractive	New
fashion designer	New	muscular	New
photographer	New	shy	New
soldier	New	social	New
fashion model	New	warm	New
		moody	New
		lazy	New
		hardworking	New
		imaginative	New
		narrow-minded	New
		boring	New
		selfish	New
		narcissistic	New

Note. “Old” groups and adjectives refer to those originally used in Pennycook et al. (2015).

3. Correlation Between LLMs' Typicality Ratings

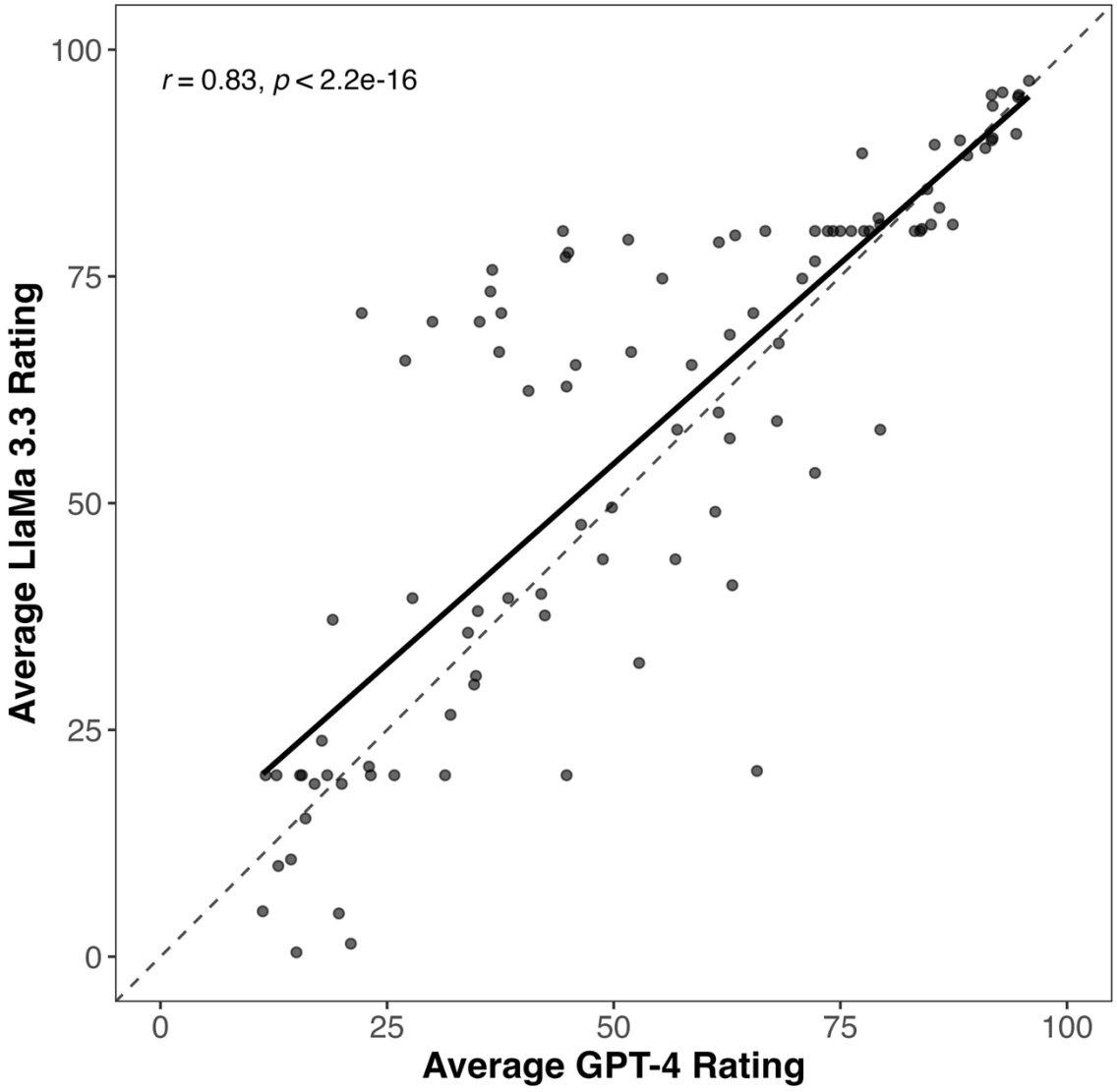

Figure S2. Relationship between the average LLM typicality ratings in Experiment 1. The solid line shows a linear model fit. Each point corresponds to an adjective–group association.

4. Validation of Prompt and LLM Settings

Table S3

Correlations between model predictions and human typicality ratings across prompt and settings variations in Experiment 1.

Model	Variation	*r*	95% CI	*p*
GPT-4	Cultural Context	0.89	[0.85, 0.93]	**< .001**
GPT-4	Current Approach	0.88	[0.83, 0.92]	**< .001**
GPT-4	Zero-shot Learning	0.88	[0.82, 0.92]	**< .001**
GPT-4	Deterministic	0.87	[0.82, 0.91]	**< .001**
GPT-4	Frequency Format	0.85	[0.79, 0.90]	**< .001**
LlaMa 3.3	Cultural Context	0.82	[0.75, 0.88]	**< .001**
LlaMa 3.3	Current Approach	0.82	[0.74, 0.87]	**< .001**
LlaMa 3.3	Deterministic	0.80	[0.72, 0.86]	**< .001**
LlaMa 3.3	Frequency Format	0.79	[0.71, 0.86]	**< .001**
LlaMa 3.3	Zero-shot Learning	0.78	[0.69, 0.85]	**< .001**

To assess the robustness of our prompt and settings, we tested four variations on the 100 group–adjective pairs from Experiment 1 to compare it with the approach we implemented throughout the paper ("Current Approach"):

- **Deterministic**: same prompt as the original one, but with temperature set to 0 to force a single deterministic output without averaging multiple responses;
- **Zero-shot learning**: removed the three examples from the user prompt;
- **Frequency format**: changed the user prompt to ask for a frequency estimation instead of a typicality rating, e.g., “Imagine a group of 100 GROUP MEMBERS. How many of them would you expect to be ADJECTIVE?”;
- **Cultural context**: specified that the stereotypes should be those prevalent in U.S. culture (since our participants were from the U.S. or Canada), by slightly changing the system prompt:

> "Your focus is to capture common societal perceptions and stereotypes *prevalent within U.S. culture*."

The results are reported in Table S3. Overall, the variation was small, and our current approach was near the best-performing settings, suggesting that our prompt and parameter choices are robust.

5. Excluding Raw Values of 0 for LlaMa 3.3 Ratings in Experiment 1

One possible explanation for the lower performance of LlaMA 3.3 compared to GPT-4 is that LlaMA 3.3 occasionally generates extreme typicality ratings of 0, which does not occur for GPT-4. To test this, we recomputed the correlation between human typicality ratings and LlaMA 3.3 in Experiment 1 after excluding all raw ratings of 0 before averaging. This exclusion did not affect the correlation ($r = 0.82$, $p < .001$). These results suggest that the lower performance of LlaMA 3.3 is unlikely to be driven by the presence of zero ratings, but rather reflects genuinely weaker model performance compared to GPT-4.

6. LlaMa 3.3 Stereotype Strength Distribution

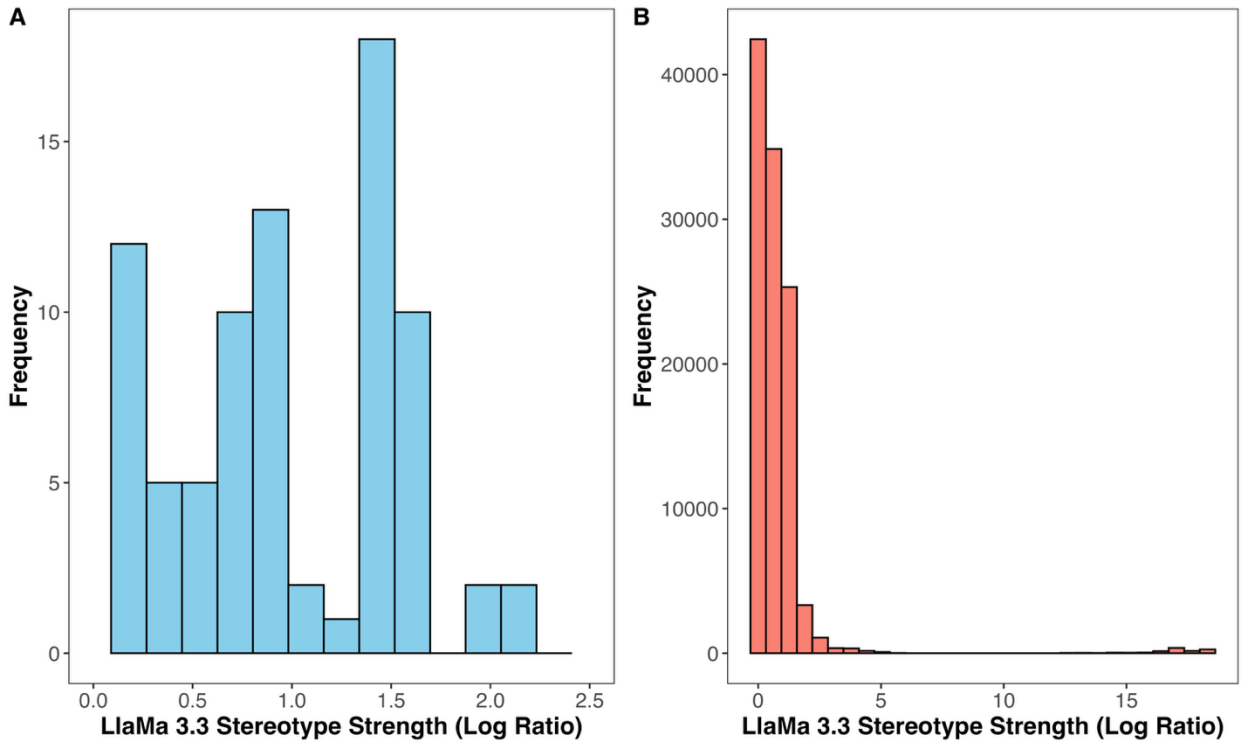

Figure S3. Stereotype strength distributions based on the LlaMa 3.3 ratings. a) Stereotype strength distribution in existing base-rate items present in our database (n = 88 items). b) Stereotype strength distribution in the new base-rate item database (n = 109,098 items). Note that the log ratios in the database are always positive by construction, since we ensured that the item with the highest typicality ratings was always in the numerator of the ratio.

7. Summary Statistics of the Database

Table S4

Summary statistics of the stereotype strength measure across all groups–adjective combinations for GPT-4 and LlaMa 3.3 in the database.

Model	*N*	Mean	Median	SD	Q1	Q3	Min	Max	Skewness
GPT-4	109,098	0.5	0.4	0.4	0.1	0.8	0	2.3	1
LlaMa 3.3	109,098	0.8	0.6	1.8	0.1	1.1	0	18.3	8

Note. N represents the number of observations. *SD* = standard deviation; *Q1* = first quartile; *Q3* = third quartile; *Min* = minimum; *Max* = maximum.

8. Example Items from the Database

Table S5

Example items from the base-rate database.

Group 1	Group 2	Adjective	GPT-4 Rating 1	GPT-4 Rating 2	GPT-4 Stereotype Strength	LlaMa 3.3 Rating 1	LlaMa 3.3 Rating 2	LlaMa 3.3 Stereotype Strength
doctor	writer	confident	88.2	72.2	0.2	86.7	73.3	0.2
chef	construction worker	idealistic	30.0	19.6	0.4	26.7	20.0	0.3
farmer	artist	helpful	80.0	38.0	0.7	80.0	40.0	0.7
executive manager	social worker	bossy	75.0	31.2	0.9	79.5	70.0	0.1
sixteen-year-old	electrician	arrogant	70.0	20.2	1.2	66.1	20.0	1.2
politician	pianist	disorganized	69.4	16.8	1.4	68.1	20.0	1.2
clown	nanny	immature	72.4	15.6	1.5	80.0	20.0	1.4
fashion designer	nanny	narcissistic	67.1	11.1	1.8	78.6	5.0	2.8
computer programmer	boxer	nerdy	85.2	10.4	2.1	83.7	5.2	2.8
fashion model	paramedic	extravagant	79.6	8.0	2.3	81.7	16.7	1.6

Note. Stereotype strength corresponds to the log ratio of the ratings as per Equation (2).