Persona-Aware Evaluation of Cognitive Bias in LLMs: From Benchmark to Applied Decision-Making

Katsumasa Yoshikawa[†‡], **Junya Takayama**[‡], **Takato Yamazaki**[‡]
[†]Dai-ichi Life Holdings, Inc., Chiyoda-ku, Tokyo, Japan
katsumasay@gmail.com

[‡]SB Intuitions Corp., Minato-ku, Tokyo, Japan
{junya.takayama, takato.yamazaki}@sbintuitions.co.jp

Abstract

We present a persona-aware evaluation suite that couples a 12-category cognitive-bias benchmark with 100 applied financial framing tasks to assess how large language models (LLMs) respond under systematically varied persona conditions. Using a factorized set of 162 personas spanning gender, age, political orientation, income, and education, we analyze how persona conditioning modulates bias-consistent responding across ten instruction-tuned models. On applied tasks, persona conditioning reduces framing reversals on average and slightly increases decision confidence, with substantial variation across model families and scales. Correlation analyses further reveal that benchmark bias tendencies—particularly availability, social proof, and framing—predict applied framing sensitivity, suggesting that standardized bias scores can serve as indicators of real-world decision variability. This work provides a unified framework for linking cognitive-bias evaluation with persona-conditioned decision behavior in LLMs. (All data and prompts will be released after acceptance to preserve anonymity.)

Keywords: LLM evaluation, cognitive bias, persona, framing, decision-making

1. Introduction

Large language models (LLMs) are increasingly deployed in settings where their outputs influence human decisions. While extensive benchmarks exist for knowledge and reasoning (Hendrycks et al., 2020; Srivastava et al., 2023; Liang et al., 2023), we still lack a systematic understanding of how LLMs behave under cognitive biases—systematic deviations from rational choice that have been documented in behavioral economics (Kahneman and Tversky, 1979; Tversky and Kahneman, 1981). Recent work has begun to show that LLMs exhibit framing-like effects (Lior et al., 2025; Germani and Spitale, 2025), raising questions about robustness and validity when such models are used in decision-support roles.

At the same time, real-world use of LLMs often involves persona prompting: explicitly specifying roles, backgrounds, or demographic traits to steer style and content. Persona conditioning can increase controllability or diversity (Li et al., 2016; Zhang et al., 2018), but it may also interact with biases in nontrivial ways. Prior studies have shown mixed effects depending on task ambiguity and context (Hu and Collier, 2024; Lutz et al., 2025), yet there has been little systematic investigation of how persona prompts modulate cognitive-bias behaviors.

Another open question is whether benchmarked bias patterns transfer to applied decision tasks. Benchmarks provide clean stress tests, but downstream use cases—such as financial choices under risk—require stable and interpretable behavior. Understanding when benchmark scores predict applied performance is essential for assessing the external validity of LLM evaluation.

This work addresses these gaps by integrating three components: (i) a unified bias benchmark covering 12 categories (120 tasks), (ii) a factorial persona space of 162 combinations, and (iii) an applied financial framing suite with matched gain/loss scenarios. Together, these resources allow us to test whether persona conditioning shifts bias-consistent responding, whether such shifts are significant, and how they relate to applied decision-making.

We address the following research questions:

- RQ1 (Bias behavior): Do LLMs display consistent, category-specific patterns on a broad cognitive-bias benchmark?
- RQ2 (Persona effect): How does persona conditioning shift decision distributions (e.g., Yes-rate) and where are these shifts statistically significant?
- RQ3 (Applied linkage): To what extent do benchmarked bias behaviors predict performance on applied financial decision-making (reversal, confidence)?
- RQ4 (Model heterogeneity): How do these effects vary across model families and sizes?

Our contributions are threefold:

1. Resource + protocol: A unified evaluation resource combining a cognitive-bias benchmark,

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4213–4225
11–16 May 2026.

a persona design, and a financial decision suite with precise metrics (Yes-rate, reversal, confidence) and statistical tests (sign, Spearman).

2. Large-scale analysis of LLMs: A systematic persona vs. no-persona comparison across bias categories and applied tasks, revealing where persona conditioning meaningfully alters decisions.

3. Linking benchmark to application: An investigation of correlations between bias-benchmark outcomes and applied financial behavior, highlighting conditions under which benchmark signals transfer—or fail to transfer—to downstream choices.

2. Related Work

2.1. LLM Benchmarks and Evaluation

General-purpose benchmarks assess knowledge, reasoning, and robustness (Hendrycks et al., 2020; Srivastava et al., 2023; Liang et al., 2023; Lin et al., 2022; Kiela et al., 2021). Beyond factuality, recent evaluations begin to examine alignment, risk, and decision behavior (Jia et al., 2024; Ouyang et al., 2025; Ross et al., 2024).

2.2. Social Bias, Fairness, and Persona-Conditioned Generation

Social-bias datasets probe fairness (Nadeem et al., 2021; Nangia et al., 2020; Parrish et al., 2022). Persona prompting has been used to control style or simulate profiles (Li et al., 2016; Zhang et al., 2018; Keskar et al., 2019; Dathathri et al., 2019), with mixed effects on factual performance and ambiguity sensitivity (Zheng et al., 2024; Hu and Collier, 2024; Lutz et al., 2025). Our study differs by pairing a factorial persona design with cognitive-bias stress tests and an applied framing task.

2.3. Cognitive Bias and Behavioral Decision-Making in AI

Behavioral economics established core effects such as prospect theory and framing (Kahneman and Tversky, 1979; Tversky and Kahneman, 1981). Recent LLM studies report human-like reframing sensitivity and source-framing biases (Lior et al., 2025; Germani and Spitale, 2025). We operationalize twelve bias categories (120 tasks) with unified Yes/No outputs and pair them with a financial framing suite to derive reversal and confidence.

```
# Persona Information
You are {persona}.

# Instruction
Always answer in JSON with the schema:
{"answer": "Yes/No", "reason": "text"}

# Scenario
A news story about a plane crash was
  widely reported this week.
Do you think air travel is now more
  dangerous than usual?

Answer:
```

Figure 1: Availability Bias Task Prompt

3. Dataset (Data Construction)

We construct three complementary resources: (i) a bias-oriented benchmark with unified binary outputs, (ii) a factorized persona space, and (iii) an applied financial decision-making suite. Because the draft submission disallows appendices, we include compact overview tables in the main paper (Tables 1–3); full templates and scripts will be released upon acceptance.

3.1. Bias Benchmark

Scope and format. We target 12 cognitive-bias categories and instantiate 10 items per category (120 total). Each item is a short vignette followed by a **Yes/No** question. *By construction, the label mapping is fixed:* ***Yes*** *is the bias-consistent answer for every item.* When drafting items, we rephrased prompts as needed so that the bias-consistent polarity aligns with YES, ensuring a uniform target across categories.

We show an example of availability bias in Figure 1. A model should generate an answer as:

```
{"answer": "Yes", "reason":
"recent vivid events feel
more frequent"}.
```

Item creation. Starting from category-wise templates, we used *GPT-4o* [1] to generate multiple candidate items per category and retained those that satisfied our normalization constraints below. Each selected item was manually reviewed and minimally edited to ensure clarity, bias-consistent labeling, and conceptual validity. This procedure increases surface diversity while keeping the decision structure comparable across items.

[1] https://platform.openai.com/docs/models

Normalization and controls. (i) Unified *binary* outputs reduce formatting variance across models; (ii) positive/negative wordings and surface forms are varied within each category to limit superficial cues; (iii) where numeric quantities appear (e.g., prices or probabilities), magnitudes are balanced across examples to avoid trivial cues or unrealistic scales.

Labeling and metric. Let $y_i \in \{\text{YES}, \text{NO}\}$ be the model output for item i. Because YES is always bias-consistent, we report the *Yes-rate* as

$$\text{YesRate} = \frac{1}{|D|} \sum_{i=1}^{|D|} [\, y_i = \text{YES}\,],$$

where $[\cdot]$ evaluates to 1 if the condition holds and 0 otherwise (macro-averaged across categories unless stated). Table 1 lists the category inventory.

Persona-aggregated reporting (weighted). For persona-conditioned evaluation, we compute a per-persona score $\bar{I}_p = \frac{1}{|D|} \sum_{i=1}^{|D|} [\, y_{i,p} = \text{YES}\,]$ and report the *weighted Yes-rate*

$$\text{YesRate}^{\text{w}} = \frac{\sum_{p \in \mathcal{P}} w(p)\, \bar{I}_p}{\sum_{p \in \mathcal{P}} w(p)}, \tag{1}$$

using the persona weights $w(p)$ defined in Eq. 2.

Output schema (JSON). Each bias item requires a strict JSON object with two fields: `"answer"` and `"reason"`. The `"answer"` must be exactly one of {YES, NO}. The free-text `"reason"` is collected for qualitative analysis and error auditing but is *not* used in the primary metric (§3 3.1).

Relation to prior datasets. Our setting differs from prior resources in three ways: (1) **Construct focus:** It targets behavioral–cognitive biases rather than social stereotypes and fairness as in StereoSet (Nadeem et al., 2021) or BBQ (Parrish et al., 2022). (2) **Unified outputs:** It enforces a unified Yes/No schema for comparability, unlike open-ended framing tasks such as WildFrame (Lior et al., 2025). (3) **Persona coupling:** It integrates persona prompting directly into bias evaluation (Li et al., 2016), enabling controlled tests of demographic modulation. Together with the applied financial suite in Section 3.3 (matched gain/loss framing), this design supports cross-space analyses of bias behaviors and their transfer to downstream decisions.

3.2. Persona Settings

Attributes. We define a balanced factorial space over five attributes: **Gender (2)** $\times$ **Age (3)** $\times$ **Politicals (3)** $\times$ **Income (3)** $\times$ **Education (3)**, yielding

Category	Example cue (short)
Anchoring	Influence of reference values
Framing	Risky vs. certain framing
Confirmation	Seek confirming evidence
Availability	Ease of recall drives frequency
Loss aversion	Losses loom larger than gains
Status quo bias	Preference for defaults
Sunk cost fallacy	Past cost drives continuation
Social proof	Follow majority behavior
Hindsight	"Knew-it-all-along" effect
Representativeness	Typicality over base rates
Overconfidence	Confidence exceeds accuracy
Endowment	Ownership inflates valuation

Table 1: Bias categories overview (compact).

Attribute	Values (with marginal weights)
Gender	male (0.5), female (0.5)
Age	25 (0.3), 40 (0.4), 60 (0.3)
Politics	liberal (0.3), moderate (0.3), conservative (0.4)
Income	low (0.3), middle (0.5), high (0.2)
Education	high school (0.5), university (0.4), graduate (0.1)

Table 2: Persona attribute design. Each attribute has discrete values; numbers in parentheses indicate the marginal weight (summing to 1 within each attribute).

162 personas. In addition, we attach two *randomized contextual fields* drawn uniformly from fixed eight-category lists: **Country** $\in$ {USA, Germany, India, Brazil, Kenya, France, China, Japan} and **Occupation** $\in$ {student, engineer, teacher, doctor, artist, farmer, manager, freelancer}. These two fields are *not* part of the factorial persona key (they do not change the count 162) and are sampled independently per prompt instance to increase ecological validity without injecting task-specific hints.

Artificial population weighting (default for benchmark aggregation). Because real populations are not uniformly distributed across attributes, we define a simple *worldwide* marginal weighting scheme (informed by public sources)(OECD Family Database, 2025; Pew Research Center, 2025). The marginals (each summing to 1) are listed in Table 2 and induce persona weights

$$w(p) = P(g)\, P(a)\, P(\text{pol})\, P(\text{inc})\, P(\text{edu}), \tag{2}$$

where $P(\cdot)$ denotes the *marginal weight* (probability) assigned to the chosen attribute value for each attribute (Gender, Age, Politics, Income, Education). We assume independence across attributes so the joint weight factorizes as in Eq. 2; normalization is handled by the denominator in Eq. 1.

We use these weights by *default* when aggregating benchmark scores across personas. Unless noted, the randomized *country* and *occupation* fields remain unweighted (uniform).

Prompting. Persona text is prepended to the user prompt (kept constant otherwise). Unless noted, we treat all 162 personas *uniformly*. Namely, during generation, all 162 personas are sampled with equal probability; population weights (Eq. 2) are applied only during aggregation. The *country* and *occupation* fields are appended as neutral context (e.g., "Country: Japan; Occupation: engineer") and are treated as non-stratified randomization variables unless otherwise specified.

Intended use. Persona conditioning enables between-condition comparisons (*persona* vs. *no-persona*) and within-condition analyses across attributes (e.g., age or education). We quantify significance with sign test in Section 5. Table 2 lists the full factorial grid of attributes.

3.3. Applied Financial Tasks

Unlike the benchmark tasks in §3.1, which serve as diagnostic probes of bias-consistent responding in short abstract vignettes, the applied suite here tests decision-level consequences of such biases in realistic gain/loss contexts.

Design and instantiation. We evaluate **100 matched gain–loss pairs** instantiated directly from a parametric *prompt* template (schema in Table 3). Each pair shares the same numeric parameters but flips wording between gain and loss frames; the decision skeleton is fixed (A=CERTAIN, B=RISKY). Amounts and probabilities are sampled from bounded grids and filtered so that the risky and certain options have comparable expected values (within a small tolerance), avoiding trivial dominance. Prompts require a strict JSON with `choice`$\in \{A, B\}$, `confidence`$\in [0, 1]$, and a short `rationale`; outputs are schema-validated (re-queried if invalid). For persona runs, the persona preface is prepended verbatim, while the decision text and parameters remain identical across persona/no-persona conditions.

A typical short example of *Gain frame* is:

> Now, you have $10,000. The market is ... Choose one of the following two options:
>
> Option A (Certain): You will gain $2,000.
> Option B (Risky): 50% chance to gain $4,000, 50% chance to gain nothing.

```
{"choice": "A", "confidence": 0.7, "rationale": "risk-averse in gains"}
```

Frame	Option A (Certain)	Option B (Risky)
Gain	e.g., sure $+\$X$	50% $+\$Y$, 50% $+\$0$
Loss	e.g., sure $-\$X$	50% $-\$Y$, 50% $-\$0$

Table 3: Applied financial task schema (paired Gain/Loss).

Metrics. For a matched frame pair p, let $c_p^{\text{gain}}, c_p^{\text{loss}} \in \{\text{CERTAIN}, \text{RISKY}\}$. The pair is a *reversal* if $c_p^{\text{gain}} \neq c_p^{\text{loss}}$; the *reversal rate* is

$$\text{RevRate} = \frac{1}{P}\sum_{p=1}^{P}[\, c_p^{\text{gain}} \neq c_p^{\text{loss}}\,].$$

We also report the *confidence mean* as the average of reported confidences (clipped to $[0, 1]$):

$$\overline{\text{Conf}} = \frac{1}{N}\sum_{i=1}^{N}\text{conf}_i \in [0, 1].$$

All metrics are computed with and without persona conditioning for all ten models. Table 3 outlines the applied-task schema.

Reproducibility. We will release full prompt templates, persona strings, item keys, and analysis scripts (Yes-rate, reversal, confidence) to facilitate reproduction of all results.

4. Experimental Setup

Models. We evaluate ten contemporary *instruction-tuned* LLMs covering three families and sizes: *Qwen* [2](Qwen2.5-7B, Qwen2.5-72B; Qwen3-8B, Qwen3-32B, Qwen3-30B-A3B), *Llama* [3](Llama-3.1-8B, Llama-3.1-70B), and *Mistral* [4](Mistral-7B, Mixtral-Large, Mixtral-8$\times$22B). All models are queried in both *no-persona* and *persona-conditioned* modes (§3).

Evaluation protocol. All models are tested in both no-persona and persona-conditioned modes using the unified bias benchmark (§3.1) and applied framing suite (§3.3). Prompts follow strict JSON formats, and invalid outputs are automatically re-queried once. Decoding is deterministic (temperature = 0) to ensure reproducibility. Persona weights (Eq. 2) are applied only during aggregation, as detailed in §3.2. All prompts, templates, and analysis scripts will be released to facilitate replication.

[2] https://huggingface.co/Qwen
[3] https://huggingface.co/meta-llama
[4] https://huggingface.co/mistralai

Model	No-persona	Persona	Δ (pp)
Qwen2.5-7B	0.175	0.095	-8.03
Qwen2.5-72B	0.200	0.261	6.11
Qwen3-30B-A3B	0.325	0.287	-3.78
Qwen3-235B-A22B	0.267	0.214	-5.26
Meta-Llama-3.1-8B	0.342	0.280	-6.16
Meta-Llama-3.1-70B	0.200	0.277	7.72
Llama-3.3-70B	0.183	0.257	7.37
Mistral-7B	0.342	0.285	-5.70
Mistral-Large	0.283	0.216	-6.75
Mixtral-8x22B	0.126	0.107	-1.87

Table 4: Benchmark Yes-rate (weighted) by model. Δ is persona $-$ no-persona in percentage points (pp).

5. Results

5.1. Benchmark Performance

Table 4 reports weighted Yes-rates on the bias benchmark, where YES is the bias-consistent label by construction (Section 3). Averaged over models, the no-persona condition attains 0.244 while persona attains 0.228 ($\Delta = -1.64$ percentage points; weighted aggregation over personas).

The effect of persona prompting is model-dependent: three models show increases in bias-consistent Yes (*Meta-Llama-3.1-70B* $+7.72$pp, *Llama-3.3-70B* $+7.37$pp, *Qwen2.5-72B* $+6.11$pp), whereas several models show decreases (*Qwen2.5-7B* -8.03pp, *Mistral-Large* -6.75pp, *Meta-Llama-3.1-8B* -6.16pp). The highest overall Yes-rate without persona is tied between *Mistral-7B* and *Meta-Llama-3.1-8B* (both 0.342), whereas with persona it is *Qwen3-30B-A3B* (0.287). These results indicate that persona conditioning can either amplify or attenuate bias-consistent responding depending on the model family and scale.

Significance. Using paired sign tests ($p<.05$) on item-level aggregates, we detected category-level significance only for a few models: *Qwen2.5-72B* (increase in 2 categories; *framing* and *confirmation*) and *Llama-3.3-70B* (increase in 1 category; *framing*), whereas *Qwen3-30B-A3B* showed a decrease in 1 category (*loss aversion*). All other models exhibited no category-level significance under the per-category $n=10$ setting. Given the small n, we treat these as indicative rather than definitive and focus on descriptive patterns below.

5.2. Persona-level Analysis

In addition to aggregate effect across models, we analyze how persona conditioning changes bias-consistent responding in detail.

Attribute level	Qwen2.5-72B	Llama-3.3-70B	Mistral-Large
gen: female	0.256	0.252	0.217
gen: male	0.266	0.262	0.215
age: 25	0.256	0.242	0.219
age: 40	0.260	0.259	0.212
age: 60	0.267	0.270	0.218
pol: conservative	0.356	0.314	0.291
pol: moderate	0.200	0.221	0.189
pol: liberal	0.195	0.218	0.143
inc: high	0.240	0.213	0.185
inc: middle	0.257	0.246	0.215
inc: low	0.282	0.305	0.237
edu:graduate school	0.227	0.207	0.160
edu:university	0.242	0.229	0.196
edu:high school	0.283	0.290	0.243

Table 5: Attribute-level slices (weighted Yes-rate; 0–1) for three cross-family models.

Attribute-level slices (cross-family view). To illustrate how persona *attributes* relate to bias-consistent responses, we report weighted Yes-rates sliced by key attributes (gender, age, politics, income, education) for three representative, cross-family models: *Qwen2.5-72B*, *Llama-3.3-70B*, and *Mistral-Large*. This choice avoids confounding architecture with scale while keeping table size manageable. The slices in Table 5 reveal model-specific sensitivities across attributes (e.g., broader spreads along some attributes than others), but consistent rankings are not universal across models; we therefore use these slices to contextualize model-level deltas.

Typical contrasts across attribute levels. Across the three cross-family models in Table 5, we observe the following regularities, computed with persona weights $w(p)$:

- **Politics** tends to show the *widest separation* among attribute levels; in all three models, conservative personas yield higher weighted Yes-rates than liberal ones. The gap is largest for Qwen2.5-72B, and slightly smaller for Llama-3.3-70B, indicating that political orientation consistently produces strong within-model contrasts even when overall Yes-rates differ across architectures.
- **Education** and **Income** exhibit *clear and systematic gradient* gradients (e.g., high school $>$ university $>$ graduate; low $>$ middle $>$ high), but these patterns are not universal across models.
- **Gender** and **Age** show only modest variation across personas. Differences are small

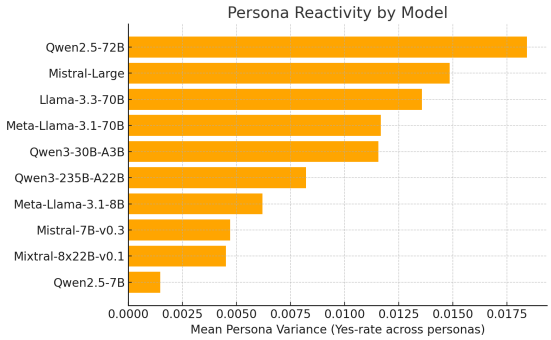


Figure 2: Persona reactivity by model. Mean variance of Yes-rates across 162 personas.

and inconsistent across models, with no clear monotonic trend by age or stable gender gap. This contrasts with findings from human behavioral studies, where gender and age often modulate risk preferences and framing sensitivity (Byrnes et al., 1999; Best and Charness, 2015).

These patterns likely reflect how *bias-mitigation* or safety tuning operates unevenly across demographic variables: attributes such as gender and age have often been subject to explicit fairness constraints, reducing variation, whereas traits like political orientation or education level are less explicitly regularized during instruction tuning, leading to greater dispersion across personas.

Persona reactivity. To quantify how strongly persona conditioning affects model behavior, we compute the variance of Yes-rates across 162 personas for each model–task pair. Figure 2 shows the mean variance per model. Larger models generally exhibit higher persona reactivity, indicating greater sensitivity to persona cues, whereas mixture-of-experts (MoE) architectures such as Mixtral or Qwen3 display milder variation—suggesting that expert routing partially absorbs persona effects. This intra-model variability complements the cross-attribute analyses above, highlighting that LLMs differ not only in average persona effects but also in the consistency of their responses under persona conditioning—an important consideration for simulations involving heterogeneous agent populations.

5.3. Applied Financial Tasks

We evaluate 100 matched gain–loss pairs per model, reporting two metrics: *reversal rate* (fraction of pairs where the choice flips between gain and loss; lower indicates more frame-invariant decisions) and the *confidence mean* (0–1, averaged over pairs). Both metrics are computed per persona and aggregated with weights $w(p)$; the no-persona baseline is computed directly on the same pairs.

Overall patterns. On average across models, persona reduces reversal rate from 36.3% to 32.3% ($\Delta = -3.96$pp) and slightly increases confidence from 80.9% to 81.7% ($\Delta = +0.86$pp). The direction, however, remains model-dependent (Table 6): about two-thirds of the models show reduced reversal—indicating more frame-invariant choices—whereas a few, notably the Qwen3 series, display large increases.

Model highlights. The lowest persona reversal is observed for *Mixtral-8×22B* (0.1%), paired with the highest persona confidence (99.5%). The largest absolute persona shift in reversal occurs for *Qwen3-30B-A3B* ($|\Delta| = 51.6$pp; increase), while *Qwen3-235B-A22B* shows the largest absolute confidence shift ($|\Delta| = 8.0$pp).

Interestingly, this contrasts with the Qwen2.5 series, where persona prompting reduces reversal (i.e., yields more frame-invariant choices), whereas the newer Qwen 3 models show increased reversal. This divergence likely reflects differences in alignment objectives and persona integration: Qwen3's stronger contextual conditioning makes it more responsive to persona framing cues, while Qwen2.5 remains closer to probabilistic consistency. Overall, persona prompting substantially modulates framing sensitivity and reported certainty, but the direction and magnitude vary across model families and tuning stages.

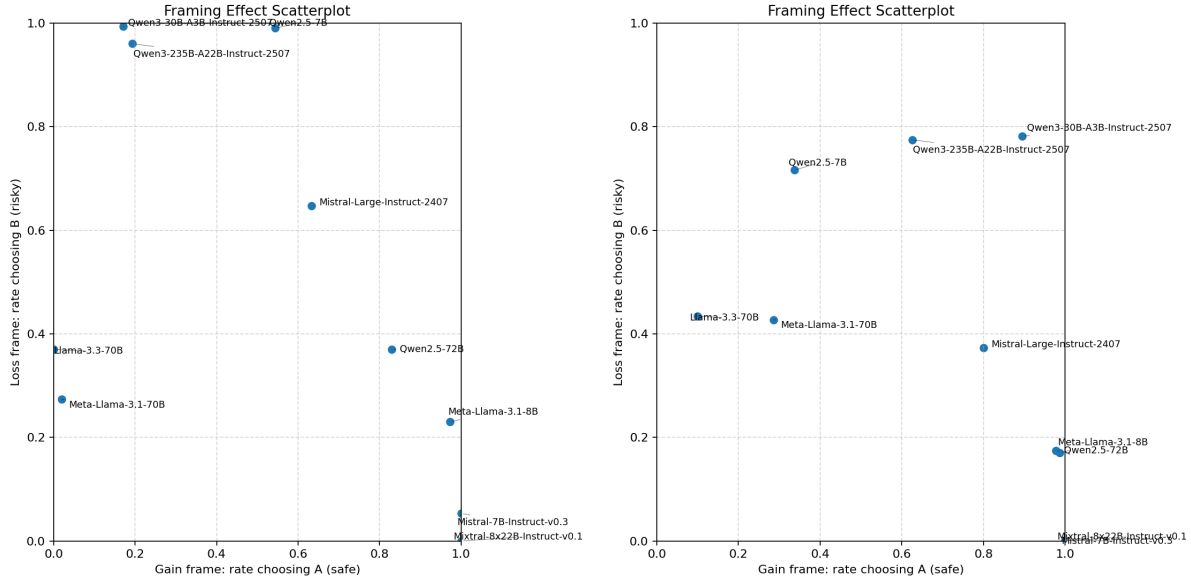
Framing-map interpretation. Figure 3 plots, for each model, the rate of choosing the *safe* option in the gain frame (x-axis) against the rate of choosing the *risky* option in the loss frame (y-axis). The *upper-right* area corresponds to the classic framing pattern (risk-averse in gains, risk-seeking in losses) and therefore indicates high reversal, whereas the *lower-right* and upper-left areas represent frame-invariant policies (consistently risk-averse or risk-seeking, respectively).

No-persona vs. persona (qualitative shifts). Comparing the two panels in Figure 3, no-persona points cluster near the edges (i.e., $x \approx 0/1$ or $y \approx 0/1$), showing near-deterministic choices that lead to high reversal. Under persona conditioning, many points move inward toward the center, suggesting attenuated heuristic saturation and more parameter-dependent behavior.

Shifts toward the lower-right or upper-left indicate greater frame-invariance and align with the average reversal drop in Table 6, whereas a few models move toward the upper-right, amplifying

Model	Reversal rate			Confidence mean		
	No-persona	Persona	Δ (pp)	No-persona	Persona	Δ (pp)
Qwen2.5-7B	54.0	14.1	-39.9	75.1	79.7	4.6
Qwen2.5-72B	47.3	19.2	-28.1	81.1	78.6	-2.5
Qwen3-30B-A3B	17.7	69.3	51.6	85.4	85.3	-0.1
Qwen3-235B-A22B	22.0	47.8	25.8	66.6	74.6	8.0
Meta-Llama-3.1-8B	24.3	17.2	-7.2	85.8	85.2	-0.6
Meta-Llama-3.1-70B	73.3	57.8	-15.6	71.4	71.0	-0.4
Llama-3.3-70B	63.0	56.5	-6.5	70.5	66.2	-4.3
Mistral-7B	5.3	0.2	-5.1	98.9	99.1	0.2
Mistral-Large	56.0	41.4	-14.6	73.7	78.1	4.4
Mixtral-8x22B	0.0	0.1	0.1	100.0	99.5	-0.5

Table 6: Applied financial tasks (100 matched gain–loss pairs): reversal rate (%) and confidence mean (%) by model. Δ is persona − no-persona in percentage points (pp). Lower reversal indicates more frame-invariant choices.



No-persona *Persona*

Figure 3: Framing map of gain–loss choices: x-axis = rate choosing A (safe) in gain; y-axis = rate choosing B (risky) in loss. Top-right = classic framing (high reversal); bottom-right/top-left = policy-stable (low reversal).

the classic framing effect. For example, *Qwen2.5-72B* drifts lower-right (reversal ↓), *Qwen3-30B-A3B* moves upper-right (reversal ↑), *Mixtral-8×22B* anchors the lower-right with near-zero reversal, and *Mistral-Large* stays near the diagonal with modest change.

Link to summary metrics. The geometry of Figure 3 thus directly explains the table-level results: proximity to the upper-right predicts higher reversal, whereas lower-right or upper-left positions indicate more frame-invariant decision policies. Persona prompting can therefore reshape a model's qualitative decision pattern, not merely its aggregate rates.

Condition	Median ρ (Yes↔Rev)	#Sig	Median ρ (Yes↔Conf)	#Sig
nopersona	-0.214	0	0.261	0
persona	0.364	3	-0.079	1

Table 7: Across-bias Spearman correlations between benchmark (weighted Yes-rate) and applied metrics. Values are median over 12 bias categories; #Sig indicates the number of categories with statistically significant correlations ($p<.05$).

5.4. Correlation: Benchmark ↔ Applied

We correlate, across models, the benchmark weighted Yes-rate (Section 3) with the applied met-

rics from §5.3—reversal rate (lower is better frame-invariance) and confidence mean. For each bias category (12 total), we compute a Spearman ρ over the k models (k=10) and summarize the distribution of ρ across categories for each condition (no-persona vs. persona).

Table 7 shows that under *persona*, benchmark Yes-rate is *positively* associated with applied reversal (median ρ=0.364, $3/12$ categories with p<.05), while the association with confidence is weakly negative (median ρ=−0.079, $1/12$ at p<.05). By contrast, under *no-persona*, the median association flips sign for reversal (median ρ=−0.214, $0/12$) and is small positive for confidence (median ρ=0.261, $0/12$).

Which categories drive the signal? Under *persona*, the strongest positive links between benchmark Yes-rates and applied *reversal* are observed for *availability* (ρ=0.903, p=3.4×10^{-4}) and *social proof* (ρ=0.779, p=7.9×10^{-3}); importantly for our setting, the benchmark *framing* category ranks *third* by $|\rho|$ with ρ=0.733 (p=1.58×10^{-2}).

These results indicate that certain bias categories—especially those related to heuristic recall or social influence—align most strongly with applied framing sensitivity across models. For *confidence*, *social proof* shows a significant negative association (ρ=−0.767, p=9.64×10^{-3}), while other categories exhibit mixed or weak patterns.

Overall, persona prompting appears to align benchmarked susceptibilities with applied framing sensitivity: models showing stronger bias-consistent responding on these categories also tend to exhibit higher reversal. Given the limited sample (k=10 models per category), these findings should be interpreted descriptively rather than as causal mechanisms.

6. Discussion

Answering the research questions. We answer the four research questions as follows:

RQ1 (*Do personas modulate benchmark bias?*) Yes. Personas modulate bias-consistent responding, but in a model-dependent manner. Some models show stronger bias under persona conditioning while others weaken, suggesting that instruction-tuning style and scale jointly determine susceptibility.

RQ2 (*Which persona attributes/categories matter?*) Attribute slices Political orientation and income show the clearest separations, whereas gender and age remain stable—likely reflecting explicit bias-mitigation on common demographic variables. These attribute slices contextualize the aggregate shifts observed across models.

RQ3 (*Effect on applied financial decisions?*) On average, personas slightly reduce framing reversals and raise confidence, indicating less heuristic saturation and more stable decision behavior, though some models amplify the classic framing pattern.

RQ4 (*Do benchmarks predict applied behavior?*) Yes, Benchmarked bias tendencies partially transfer to applied framing sensitivity: higher bias-consistent Yes-rates correlate with stronger reversal, especially in categories such as availability, social proof, and framing itself. Overall, persona conditioning alters both benchmark and applied decision profiles in a model- and attribute-dependent manner, providing a systematic link between bias evaluation and real-world decision contexts.

Implications for evaluation and use. Persona prompts are not uniformly "helpful" or "harmful"; they can shift decision regimes, either stabilizing responses (consistently risk-averse) or amplifying classic framing.

Because such shifts are model- and attribute-dependent, evaluation reports should include both persona and no-persona conditions, tracking (i) aggregate deltas, (ii) framing-map locations, and (iii) key attribute slices to reveal model–persona interactions.

Attributes such as political orientation and income often show the largest separations and are therefore informative for fairness or interpretability audits.

When decision stability is required, models that occupy the lower-right (risk-averse) or upper-left (risk-seeking) regions of the framing map are preferable to those near the upper-right (high reversal under classic framing).

Where next? Future work should expand the item pool and domains (e.g., health or legal framing), extend to multilingual settings and richer persona attributes, and explore alternative weighting schemes.

Causal analyses could test which persona tokens drive shifts on the framing map and evaluate interventions—such as fine-tuning or guardrails—that preserve task accuracy while reducing undesired framing sensitivity. All data, prompts, and analysis scripts will be released to facilitate replication and meta-analysis.

7. Conclusion

We introduced a persona-aware evaluation suite coupling a 12-category cognitive-bias benchmark with 100 applied financial framing tasks, and evaluated ten contemporary LLMs. Persona conditioning was found to modulate bias-consistent responding in a model-dependent manner; on applied tasks, it

reduced reversal on average and slightly increased confidence, with variation across model families and scales.

Benchmark Yes-rates under personas correlated with applied framing sensitivity—especially for availability, social proof, and framing—indicating that benchmark scores can serve as informative predictors of downstream decision behavior. Together, these results provide a unified lens for linking bias evaluation to applied decision-making, offering guidance for future model auditing and socially grounded LLM simulations.

Ethical Considerations and Limitations

Ethical considerations. Our personas encode sociodemographic attributes (gender, age, politics, income, education, plus country/occupation). Although we avoid derogatory content, persona prompts may *prime stereotypes* or shift model behavior in ways that affect downstream users. We aggregate with coarse artificial weights to approximate global marginals, which are not substitutes for locale-specific distributions. We release prompts, scripts, and aggregated results; no personal data are collected.

Limitations. Per-category sample size is small (n=10), limiting statistical power; we therefore emphasize aggregate deltas and descriptive patterns. Weights are simplified (global, static); real deployments may warrant regional/sector-specific marginals. Personas and bias items were LLM-generated (GPT-4o), which may imprint style artifacts; prompts are English-only; decoding/system settings can matter. Applied tasks cover a single domain (financial framing) and one confidence format. Correlations are computed across k=10 models and should be read as descriptive, not causal.

8. Bibliographical References

Ryan Best and Neil Charness. 2015. Age differences in the effect of framing on risky choice: A meta-analysis. *Psychology and Aging*, 30:688–698.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

James Byrnes, David Miller, and William Schafer. 1999. Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, 125:367–383.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *CoRR*, abs/1912.02164.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Federico Germani and Giovanni Spitale. 2025. Source framing triggers systematic evaluation bias in large language models.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E. McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for llms under uncertain context.

Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Gili Lior, Liron Nacchace, and Gabriel Stanovsky. 2025. Wildframe: Comparing framing in humans and llms on naturally occurring texts.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

OECD Family Database. 2025. Sf1.4: Population by age of children and young adults, and youth dependency ratio. Indicator SF1.4.

Shumiao Ouyang, Hayong Yun, and Xingjian Zheng. 2025. Ai as decision-maker: Ethics and risk preferences of llms.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Pew Research Center. 2025. Methodology: Ideology classification (left, center, right).

Jillian Ross, Yoon Kim, and Andrew W. Lo. 2024. Llm economicus? mapping the behavioral biases of llms via utility theory.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari,

Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with*

books, 20th edition. The Phantom Editors Associates, Gotham City.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models.