Unveiling Latent Cognitive Bias Constructs in Large Language Models

Bezawit Lake Tilaye
School of Computer Science
University of Bristol
dz24190@bristol.ac.uk

Huw Day
School of Engineering Maths
& Technology
University of Bristol

Chris McWilliams
School of Engineering Maths
& Technology
University of Bristol

Telmo M. Silva Filho
School of Engineering Maths
& Technology
University of Bristol

Abstract

As Large Language Models (LLMs) are increasingly adopted in decision-making, there is a growing need to better understand the behavioural patterns underlying their reasoning. Evaluation is one of the primary approaches used to study model behaviour. However, most existing LLM evaluations rely on task-based benchmarks that summarise performance through aggregate scores. While effective for comparing model performance, such approaches provide limited insight into the latent behavioural tendencies that drive biased reasoning. In this paper, we propose a construct-based evaluation framework for studying cognitive bias in LLMs grounded in psychometric evaluation. We systematically perturb prompts across controlled conditions to analyse reasoning behaviour. From model responses, we derive behavioural features capturing susceptibility, robustness, correction, and variability, and apply latent variable modelling to infer underlying behavioural constructs. Our findings show that cognitive bias in LLMs is structured and dynamic, varying across bias type, model families, and scale.

1 Introduction

Large Language Models (LLMs) demonstrate strong performance across language understanding, reasoning, and problem-solving tasks (Pan et al., 2024; Dziri et al., 2023). However, their internal decision-making mechanisms remain poorly understood, making it difficult to evaluate how these systems reason, generalize, and make decisions reliably (Tamkin et al., 2021; Burnell et al., 2023).

Humans are boundedly rational and systematically biased in their judgments and decision-making processes (Tversky and Kahneman, 1974). These lead to systematic deviations from rational reasoning, characterized as **cognitive bias (CB)** (Gilovich et al., 2002). LLMs are trained on human-generated data and further fine-tuned using human instructions and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). As a result, they are likely to inherit not only linguistic structure but also statistical regularities of human reasoning, including CBs embedded in training data and feedback signals (Caliskan et al., 2017).

Existing discussions of reliability and fairness in LLMs have focused primarily on societal biases, including gender, political, and ethical biases (Wan et al., 2023; Bowen III et al., 2024; Rozado, 2024). These biases often stem from social and historical inequalities in training data, manifesting as discriminatory behaviour toward demographic groups (Wan et al., 2023). In contrast, CBs emerge from systematic error tendencies in reasoning and judgment itself, often through heuristic decision-making processes under uncertainty (Tversky and Kahneman, 1974). As a result, CBs are expressed less as explicit discriminatory outputs and more as behavioural tendencies, making them inherently more difficult to identify and evaluate.

CB evaluations for LLMs typically rely on task-based benchmarks (Burden et al., 2025), for example, a model may exhibit 45% confirmation bias and 25% anchoring bias, these scores capture the frequency of biased responses but not the shared mechanisms and behavioural tendencies underlying them (Burnell et al., 2023). While behaviour may vary across contexts and prompt formulations, task-based evaluations provide limited insight into whether such variation reflects transient prompt effects or consistent reasoning patterns.

This limitation highlights a mismatch between the nature of CB and evaluation paradigms often used to measure it. CBs are not isolated prediction errors or task-specific failures; rather they reflect underlying tendencies in reasoning and judgment manifesting across contexts and decision-making conditions (Tversky and Kahneman, 1971). As shown in Figure 1 task-based evaluations, instead, treat response as independent observations reduc-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37438–37458
October 24-29, 2026 ©2026 Association for Computational Linguistics

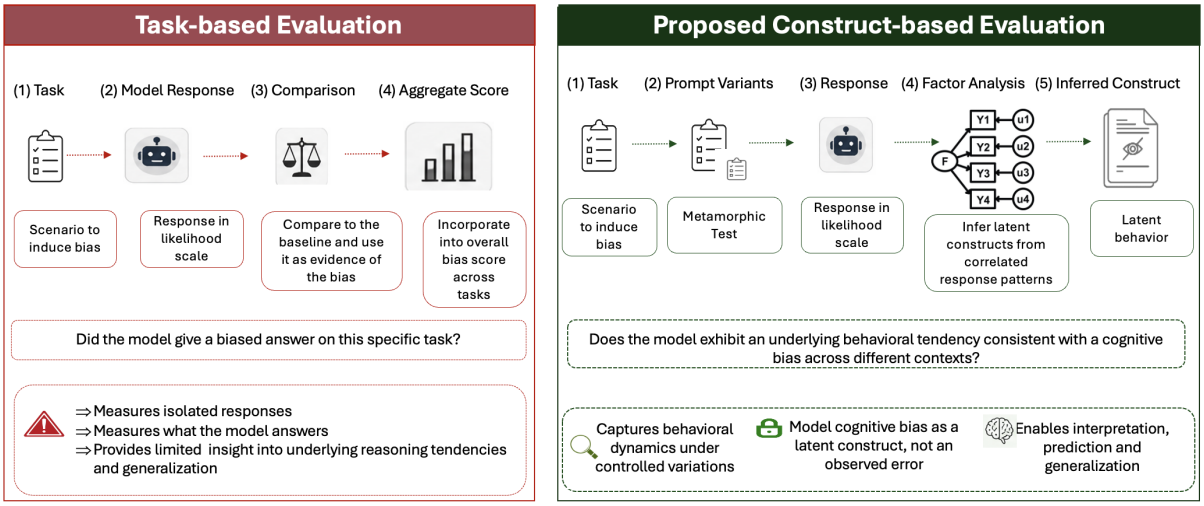


Figure 1: Task-based evaluation and the proposed construct-based evaluation framework for studying CBs in LLMs. Task-based approaches evaluate isolated prompt-response outcomes and aggregate them into benchmark bias scores, whereas the proposed framework models CBs as a latent behavioural construct inferred from correlated response patterns across systematically perturbed prompting conditions.

ing behaviour to aggregate metrics, obscuring the latent structure that gives rise to biased reasoning (Burden et al., 2025).

Psychometrics provides a framework to address this limitation (Bowen III et al., 2024) by treating unobservable psychological attributes, such as intelligence and reasoning ability, as latent constructs from observed behavioural patterns (Burnell et al., 2023). Similarly, CBs reflect underlying tendencies that cannot be directly measured from single-task outcomes. Thus, in this paper we propose a construct-based framework to study CBs in LLMs.

We model CBs as latent behavioural constructs emerging across prompting conditions and decision-making contexts. To assess these patterns, we systematically perturb prompts and analyse how model responses change under these transformations. We then use latent variable modelling to infer CB constructs from correlated response patterns across conditions, going beyond aggregate scores toward interpretable and generalizable accounts of CB in LLMs. Our contributions are: (i) a construct-based framework for evaluating CBs in LLMs grounded in psychometrics; (ii) a structured perturbation methodology that systematically modifies prompts across multiple conditions to analyse behavioural sensitivity and stability; and (iii) latent variable modelling of CBs as emergent behavioural constructs inferred from correlated response patterns rather than isolated benchmark outcomes, leading to deep insights into CBs in LLMs.

2 Related work

LLM Benchmarks. LLM evaluation often focusses on factual knowledge (Hendrycks et al., 2020; Srivastava et al., 2023; and Phan et al., 2026), with benchmarks presenting a wide range of curated questions on a variety of topics across fields to test a model's world knowledge and problem-solving capacity. To avoid memorisation (Ahmed et al., 2026), researchers have been developing benchmarks designed to "catch out" models relying on these habits. TruthfulQA comprised questions that humans could answer falsely due to false beliefs or misconceptions (Lin et al., 2022). Top-performing models had to avoid generating false answers learned from human texts. Previous work (Liang et al., 2023) moved beyond raw knowledge/recall and binary correct/incorrect measurement, by evaluating uncertainty calibration, robustness (to typos/formatting changes etc.), fairness (consistency across demographic contexts), bias (degree of stereotypical associations), toxicity (likelihood of harmful outputs) and efficiency. Recent work (Gostev, 2026) measures whether models detect nonsense, call it out, and avoid confidently continuing with invalid assumptions.

CB and Heuristic Reasoning. Human decision-making uses heuristics, which act as cognitive shortcuts that simplify reasoning under uncertainty (Tversky and Kahneman, 1974), enabling efficient judgment, but leading to systematic deviations from normative reasoning if applied inappropri-

ately (Srivastava et al., 2023). These deviations characterize CB and commonly emerge from uncertainty, incomplete information, or probabilistic reasoning (Tversky and Kahneman, 1981). Unlike random errors, CBs are stable distortions in information perception, processing, and integration for decision-making (Coda-Forno et al., 2024).

CBs have been extensively studied in psychology and behavioural economics, leading to the identification of reasoning distortions associated with distinct heuristic mechanisms (Kahneman, 2011). In this work, we examine a representative set of biases spanning probabilistic reasoning (Base-rate Fallacy, Gambler's Fallacy), belief updating (Confirmation Bias, Hindsight Bias), affective weighting (Negativity Bias), and social influence effects (Bandwagon Effect, Suggestibility) (Haselton et al., 2015). Table 1 summarizes the selected biases and their definitions.

Bias	Definition
Anchoring	Over-reliance on initial reference values
Bandwagon Effect	Tendency to align with majority opinion or popular belief
Base-rate Fallacy	Neglect of base rates in favour of specific information
Confirmation Bias	Tendency to provide greater importance to evidence that fit existing beliefs
Gambler's Fallacy	Belief that past events affect independent future outcomes
Framing Effect	Decisions influenced by how information is presented
Negativity Bias	Greater weight assigned to negative information
Hindsight Bias	Perceiving past events as more predictable after they occur
Outcome Bias	Judging decisions based on outcomes
Suggestibility	Susceptibility to leading information
Von Restorff Effect	Improved recall for distinctive or salient items

Table 1: The set of CBs used in this study. Full definitions provided in Appendix A.

CB Evaluation in LLMs. Several benchmarks show that LLMs exhibit systematic deviations from normative decision-making behaviour, including susceptibility to a range of CBs (Malberg et al., 2024; Echterhoff et al., 2024; Shaikh et al., 2024). Coda-Forno et al. (2024) adapted classical cognitive psychology experiments to assess probabilistic reasoning and risk-related behaviour in LLMs, while also analysing the effects of model scale and reinforcement learning from human feedback (RLHF). Echterhoff et al. (2024) proposed a large-scale evaluation suite designed to probe prompt-induced and sequential CB, as well as mechanisms for mitigation and self-correction. Malberg et al. (2024) conducted an extensive analysis of 30 CBs demonstrating that biased behaviour is widespread across architectures and model families. Complementing this, Shaikh et al. (2024) used influence-based analysis to identify which prompt components are most responsible for eliciting biased reasoning patterns in frontier models.

3 Dataset Generation

Metamorphic testing (MT) is commonly used in software engineering to validate programs in the absence of test oracles for checking correctness. MT relies on metamorphic relations, expected relationships between multiple inputs and their corresponding outputs (Segura et al., 2016). For LLMs, MT assesses correctness through consistency constraints, i.e., if two inputs differ only in aspects that should not affect the underlying reasoning, the outputs are expected to remain invariant, with violations interpreted as deviations in model behaviour (Ribeiro et al., 2020). Using this framework, we systematically generate different types of prompts to probe specific aspects of heuristic reasoning, operationalized along four dimensions: (i) bias susceptibility, (ii) stability under perturbation, (iii) bias dynamics, and (iv) meta-cognitive correction, which are tested through the following metamorphic prompt types:

Minimum Functionality Test (mft): baseline decision task without bias-inducing cues, representing the reference judgment.

Cognitive Bias Test (bias): introduces a heuristic cue designed to induce a specific bias. Deviations from the mft response indicate bias susceptibility.

Invariance Tests (INV): evaluate robustness of biased judgments under irrelevant perturbations. Four variants are used: contextual distractors (INV), irrelevant temporal details (INV_{TEMP}), metadata perturbations (INV_{META}), and unrelated numerical values (INV_{NUM}).

Directional Strength Tests (DIR): manipulate

bias intensity using strengthened (DIR_{high}) and weakened (DIR_{low}) biasing signals to evaluate whether bias can be amplified or attenuated through controlled cue variation.

Opposing Test (OPP): reverses the biasing signal while preserving task structure to evaluate directional symmetry in model responses.

Meta-cognitive Test (m_{cog}): explicitly warns about potential CBs to assess self-correction.

Each baseline task is expanded into multiple prompt variants using MT, implemented as minimal localized modifications to the original scenario, allowing behavioural differences to be attributed to the intended metamorphic manipulation.

The dataset contains 11 CBs, each with 45 baseline task. Each task requires probabilistic judgment under uncertainty, where CBs are most likely to emerge. Table 2 summarizes the distribution of test conditions across bias groups. The biases were grouped into Group A and Group B based on the structural nature of the heuristic cues associated with each bias. Group A contains biases with continuous or scalar cues (e.g numerical values in Anchoring Bias or the severity of phrasing in Negativity Bias). Since these cues exist on a continuum, they can be systematically amplified or weakened, making them suitable for Directional Strength Tests (DIR). In contrast, Group B contains biases with binary or categorical cues (Hindsight Bias or Outcome Bias). Since these cues represent discrete states (whether an event occurred or not), they cannot be meaningfully scaled in strength but can be logically reversed. Therefore, Group B biases are evaluated using Opposing Tests (OPP).

Component	Details
CBs	11
Base Tasks per Bias	45
Total Base Tasks	495
Test Conditions by Group	
All (11 biases)	7 tests (CF only)
Group A (6 biases)	9 tests (CF + DIR)
Group B (3 biases)	8 tests (CF + OPP)
Total	
Total Prompt Variants	10,935
Response Scale	1–7 (Probability Bins)

Table 2: Dataset summary and distribution of MT conditions across CB groups. CF denotes common test conditions applied to all biases, including mft, bias, INV, and m_{cog}. Different test groups are assigned based on the conceptual characteristics of each CB.

All biases include a common set of test conditions, while additional tests are applied where conceptually appropriate. All prompts are evaluated zero-shot and each query is issued independently with no conversational context. The order of prompt conditions is randomized to mitigate ordering effects. We set temperature to zero to ensure reproducibility and eliminate sampling variability. Prompt templates are provided in Appendix C.

4 Feature Engineering

We calculate the following features:

Bias Susceptibility measures how much a heuristic cue shifts the model response relative to the baseline condition: Bias Susceptibility $= \frac{\text{bias}-\text{mft}}{6}$. Normalization by the maximum possible shift on the 1-7 scale constrains the feature to $[-1, 1]$, where the sign indicates direction and the magnitude reflects bias strength.

Distraction Magnitude: measures sensitivity to irrelevant perturbations under invariance conditions. For each invariance variant, it is computed as the absolute deviation from the bias condition. Larger deviations indicate that irrelevant information has altered the biased state, reflecting instability.

Correction Magnitude quantifies the reduction in bias following meta-cognitive intervention: $MCOG_{Magnitude} = |bias - mft| - |m_{cog} - mft|$.

Correction Direction is the direction of residual deviation after meta-cognitive intervention, computed as the difference between m_{cog} and mft.

Variability measures response consistency across all prompt variants for a given task and is computed as the standard deviation across conditions.

Elasticity is the magnitude of volatility under strength manipulation, measuring response shift when heuristic cues are altered, regardless of direction. For Group A, it is computed as the combined shift between strengthened and weakened conditions relative to the BIAS state: Elasticity $= |DIR_{high} - bias| + |bias - DIR_{low}|$. For Group B, it is the absolute difference between the bias and OPP conditions, capturing the response shift under bias reversal. Higher values indicate greater sensitivity to prompt perturbations and lower values suggest more rigid behaviour.

Asymmetry evaluates sensitiveness to strengthening or weakening of the bias signal (Group A only): Asymmetry $= |DIR_{high} - bias| - |bias -$

$DIR_{low}|$. Positive values indicate higher sensitivity to bias amplification, negative values indicate higher sensitivity to reduction, and near-zero values show symmetric responsiveness.

Directional Asymmetry evaluates symmetry between bias-inducing and bias-reversing conditions relative to the baseline: Directional Asymmetry $= |bias - mft| - |OPP - mft|$. Positive values indicate stronger deviation under the biased condition, negative values indicate stronger influence of the opposing condition, and values near zero indicate symmetric sensitivity.

5 Latent Variable Modelling

We use latent variables to uncover and validate hidden behavioural constructs underlying CBs in LLMs. The analysis follows three stages. **Exploratory Factor Analysis (EFA)** is conducted separately for each group to discover latent structures in behavioural responses without imposing prior assumptions (Costello and Osborne, 2005; Fabrigar et al., 1999). **Confirmatory Factor Analysis (CFA)** validates whether the identified latent constructs explain observed correlations in behaviour (Brown, 2015). Finally, **Structural Equation Modelling (SEM)** is applied to model relationships among validated constructs and observed behaviours within a unified framework (Kline, 2005).

6 Results

We evaluate 22 LLMs spanning multiple families, sizes, and alignment strategies (see Table 3).

6.1 Feature Analysis

Feature distributions show sufficient variability for subsequent analysis (see Appendix B). Sampling adequacy was evaluated using the Kaiser-Meyer-Olkin (KMO) measure. The global feature set achieved a KMO of 0.73, while Group A and Group B obtained KMO values of 0.698 and 0.623, respectively. Although the Group B value is comparatively lower, it remains within the acceptable range for EFA. Bartlett's test of sphericity was highly significant across all feature groups ($p < 0.001$), rejecting the null hypothesis that the correlation matrices are identity matrices.

Latent Factor Selection: Groups A and B both exhibited clear three-factor solutions, with eigenvalues exceeding the Kaiser criterion (> 1) and corresponding random eigenvalues. In contrast, the shared feature set supported a two-factor solution, as the third eigenvalue only marginally exceeded the random baseline (see Appendix B).

6.2 Factor Analysis

We applied EFA using the features calculated to extract latent behavioural dimensions across all groups. This was implemented using the Python package factor_analyzer with the default principal axis factoring method and Varimax rotation, which produces orthogonal factors that maximize interpretability. Factor loadings above 0.4 were considered significant. We assign interpretable names to the detected factors based on the features that loaded strongly on each of them. Three consistent factors emerged across groups (Figure 2):

Factor 1 is dominated by invariance-based features captures the extent to which models maintain stable reasoning in the presence of distractors and is interpreted as *Attentional Filtering (AF)*. **Factor 2** is associated with Bias Susceptibility and Meta-cognitive Direction. These factor reflects the directional tendency of models to follow heuristic cues and is interpreted as *Heuristic Reliance (HR)*. **Factor 3** is characterized by Bias Elasticity, Asymmetry, and variability, captures how biases evolve under manipulation and is interpreted as *Bias Dynamics (BD)*. The cumulative explained variance is 41.5% for the shared feature set, 52.3% for Group A, and 55.6% for Group B. These results suggest that the extracted factors provide a meaningful summary of the underlying behavioural tendencies across models and conditions.

6.3 Bias Factor Score

Interpretation of latent factors follows from feature construction and z-score normalization. AF is derived from absolute deviation measures, positive values indicate distractibility, while negative values reflect robustness. HR measures the magnitude and direction of heuristic influence, positive and negative values indicating upward (push) and downward (pull) shifts relative to baseline, respectively. BD characterizes behavioural flexibility, positive values indicate volatile, amplification-prone biases, whereas negative values reflect more rigid biases. Near-zero scores correspond to approximately average behaviour across models. These interpretations align with the raw response trajectories. For example, in Figure 3, the Outcome Bias condition shifts below MFT, producing negative HR, indicating movement opposite to the heuristic direction.

The AF dimension: Base-rate Neglect (-0.26),

Family	Model	Architecture	Total Params	Context	Alignment
OpenAI	GPT-3.5-Turbo	175B	—	16K	RLHF
	GPT-4	—	>1T	128K	RLHF
	GPT-5.4	MoE	—	1M	RLHF/RLAIF
Anthropic	Claude Sonnet 4.5	—	—	200K	RLHF
	Claude Opus 4.5	—	—	200K	RLHF/RLAIF
	Claude Sonnet 4.6	—	—	1M	CAI/RLHF/RLAIF
	Claude Opus 4.6	—	—	1M	CAI/RLHF/RLAIF
Google	Gemma 2 9B	Dense	9B	8K	SFT/RLHF
	Gemma 3 12B	Dense	12B	128K	SFT/RLHF
	Gemma 3 27B	Dense	27B	128K	SFT/RLHF
	Gemma 3n e4B	MoE	10B	128K	SFT/RLHF
Meta	LLaMA 3.1 8B	Dense	8B	128K	SFT/RLHF
	LLaMA 3.2 3B	Dense	3B	128K	SFT/RLHF
	LLaMA 3.3 70B	Dense	70B	128K	SFT/RLHF
Mistral	Mistral 3B	Dense	3B	131K	SFT/RLHF
	Mistral 8B	Dense	8B	262K	SFT/RLHF
	Mistral 14B	Dense	14B	262K	SFT/RLHF
Qwen	Qwen 2.5 72B	Dense	72.7B	131K	SFT/RLHF
	Qwen3 235B	MoE	235B	262K	RL
	Qwen 3.5 27B	Dense	27B	262K	—
DeepSeek	DeepSeek V3.1	MoE	671B	128K	—
	DeepSeek V3.2	MoE	671B	131K	—

Table 3: Overview of the evaluated language models, including model family, architecture type, parameter scale, context window size, and alignment strategy. A dash (–) denotes information that is unavailable or undisclosed.

Negativity Bias (-0.25), and Suggestibility (-0.23) exhibit strong robustness (negative AF), indicating that once triggered, these biases are resistant to perturbation and remain stable despite irrelevant cues. In contrast, the Bandwagon Effect ($+0.28$) and Gambler's Fallacy ($+0.22$) are more contextually fragile, showing higher distractibility and greater dependence on external cues (Figure 4).

The HR dimension: Hindsight ($+0.97$), Negativity ($+0.25$), and Base-rate Neglect ($+0.21$) push the model judgment upward. Conversely, Outcome Bias (-0.91), Framing Effect (-0.72), and Gambler's Fallacy (-0.35) induce downward shifts relative to the baseline. Bandwagon Effect and Suggestibility remain near neutral, indicating that while they may cause deviation, they lack strong, uniform directional momentum (see Appendix D).

The BD dimension: captures how reasoning error scale with heuristic intensity. Hindsight ($+0.62$), Negativity ($+0.85$), and Base-rate Neglect ($+0.58$) are highly dynamic, with stronger cues amplifying deviations in model reasoning. In contrast, Suggestibility (-0.96), Gambler's Fallacy (-0.48), Confirmation Bias (-0.39), and Anchoring (-0.18) are more rigid, showing relatively constant error magnitude regardless of cue strength, indicating limited sensitivity to manipulation.

These show that CBs in LLMs are structurally distinct reasoning tendencies, not uniform, static distortions. Some act as stable, rigid priors (Anchoring), others as context-sensitive shifts (Bandwagon), and others as dynamically amplifiable states (Hindsight). Also it demonstrates tendencies that appear negligible under weak manipulations often become highly pronounced under increased heuristic pressure, indicating that reasoning errors in LLMs must be evaluated as dynamic, scalable vulnerabilities rather than binary failure modes.

6.4 Model Factor Score

Group A tends to elicit relatively stable and symmetric responses, primarily differentiating models in terms of AF. Claude models exhibit strong stability and symmetry (AF ≈ -0.1, HR ≈ -0.1, BD ≈ -0.1 to -0.2), while Qwen models cluster near zero for HR and BD, indicating balanced responses. OpenAI models transition from earlier pull-dominant and volatile models (HR ≈ -0.3,

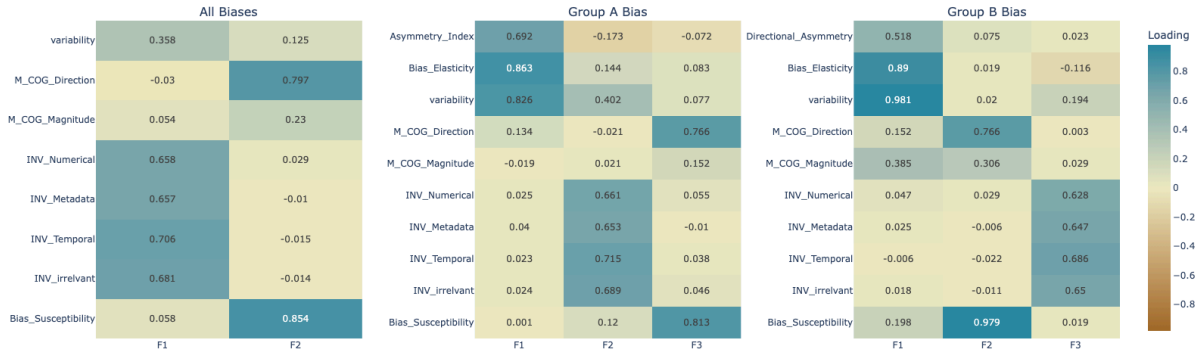


Figure 2: Factor loadings for all bias with shared features (left), Group A (DIR, middle), and Group B (OPP, right). Across all groups, three consistent factors emerge corresponding to AF, HR, and BD.

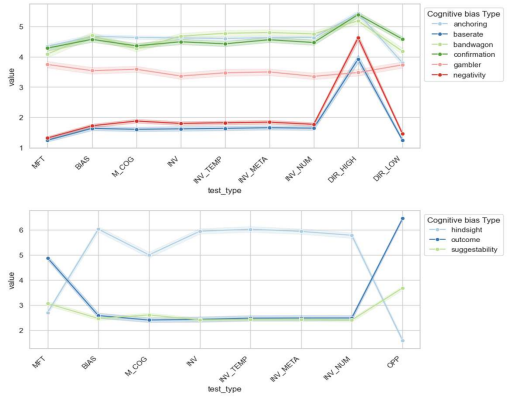


Figure 3: Raw model responses across MT conditions for Group A (top) and Group B (bottom) biases.

BD $\approx +0.2$ in `gpt-3.5-turbo`) to more neutral and stable dynamics in newer models (HR ≈ 0, BD ≈ -0.1 to -0.2). Meta-Llama shows scale dependence: larger models are more robust (AF ≈ -0.2) and shift toward push-oriented momentum (HR $\approx +0.2$). See details in Appendix E.

In contrast, Group B produces stronger directional effects and significantly higher BD, indicating that retrospective, outcome-oriented, and socially framed manipulations more effectively reshape established reasoning states. Claude models remain robust to distractors (AF ≈ -0.1 to -0.3) but shift toward moderate push-oriented influence (HR $\approx +0.1$ to $+0.2$) while maintaining relatively rigid or symmetric dynamics (BD ≤ 0). OpenAI frontier models retain mild pull-oriented momentum (HR ≈ -0.15) but exhibit substantially increased volatility (BD $\approx +0.3$), suggesting heightened amplification sensitivity. Gemma models remain comparatively stable and symmetric, with near-zero directional momentum (HR ≈ 0).

Meta-Llama again demonstrates strong scaling effects: smaller models are rigid and push-oriented (HR $\approx +0.3$, BD ≈ -1.0), whereas larger variants become more robust (AF ≈ -0.2) but shift toward pull-oriented momentum and increased volatility (HR ≈ -0.2, BD $\approx +0.4$). Qwen models display a clear scaling-driven transition from highly distractible and volatile behaviour in smaller variants (AF $\approx +0.5$, BD $\approx +0.4$) to near-symmetric and robust in larger models (AF ≈ -0.1, BD ≈ 0).

The factor analysis reveals that model families differ not only in overall susceptibility to heuristic manipulation, but also in the directional structure and dynamic evolution of induced bias states. These behaviours remain partially stable within model families except for scaling-related effects.

6.5 Structural Equation modelling

We conducted CFA and SEM for Group B, which showed the smallest sampling adequacy (KMO = 0.623) among the bias groups (Appendix B). The hypothesized model consisted of three latent constructs *AF, HR, BD* with observed indicators assigned according to the EFA structure. An acceptable overall model fit supported the validity of the proposed latent structure in explaining the observed covariance patterns.

Factor loadings support the three factor solution, with each factor aligning with its hypothesised indicators (see Appendix F). SEM reveals interactions between constructs: HR has a strong effect on BD (path coefficients ranging from 0.912 to 1.336), thus higher HR associates with increasingly volatile and dynamically amplifiable bias states. In

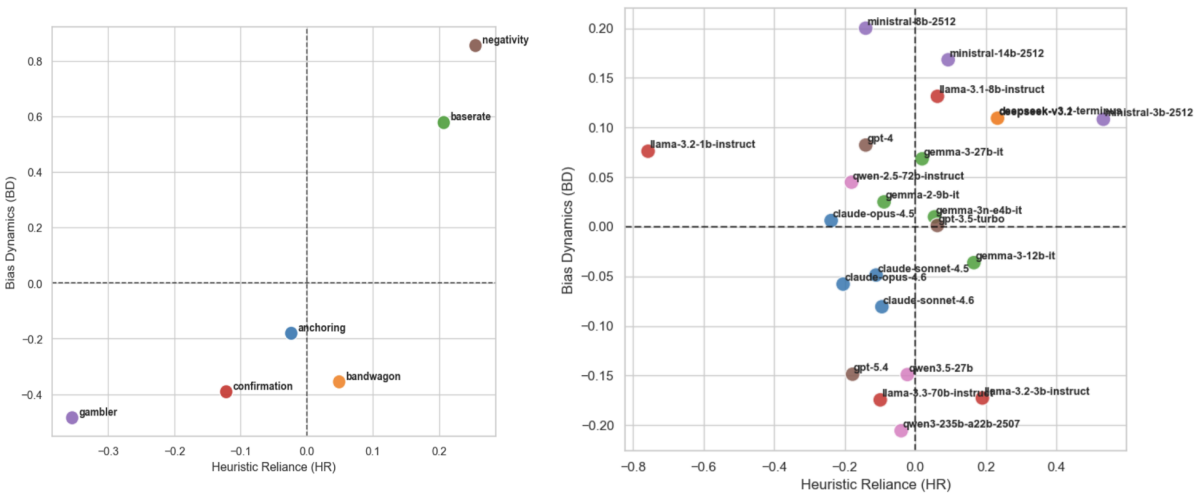


Figure 4: HR, BD across Group A CBs, illustrating model behavioural profiles under each bias condition.

contrast, AF has a modest effect on BD (-0.195): stronger AF partially mitigates BD instability.

7 Discussion

LLM CBs and Human Decision-making. CBs are often characterised by their underlying mechanisms (Gilovich et al., 2002). CBs driven by *salience and emotional weighting* reflect the human tendency to overweight vivid, recent, or emotionally charged information, leading to context-dependent shifts in judgment (Tversky and Kahneman, 1974). This is analogue to **BD latent behaviour**. CBs like Hindsight and Negativity exhibit *highly dynamic* cue intensity, indicating that model responses vary systematically with the strength of heuristic signals. When the biasing cue is strengthened, the model *actively leans further into the error, producing larger deviations*, simultaneously *shifting judgment upward* (as observed in **HR**). This suggests these CBs are not fixed responses, but scalable behavioural effects that depend on the magnitude of contextual manipulation.

Biases driven by *evaluation framing*, where judgment is influenced by the presentation or result of an event rather than the decision process itself, show human sensitivity to negative contexts (Baron and Hershey, 1988). Framing Effect and Outcome Bias induce *inverse shifts relative to the baseline* (**HR**), consistent with human responses to negative framing. Also, in Outcome Bias, where the initial negative outcome is subsequently contrasted with a positive opposing cue, the model also exhibits flexible adjustment (**BD**), showing that its reasoning state can adapt when the contextual cues change. This shows LLMs do not just amplify positive or salient cues, they mimic human biases by shifting their judgments according to the shortcut.

Belief perseverance is characterised in human decision-making as the tendency to maintain initial beliefs despite exposure to new information (Kahneman, 2011). In LLMs, a related pattern is observed where certain biases exhibit relatively stable response magnitudes across prompt types. Biases such as Suggestibility, Confirmation, and Anchoring show that, regardless of how the biasing cue is presented in the prompt, the *error magnitude remains stable*. This shows they are less sensitive to changes in cue strength, resulting in relatively invariant deviations from the baseline. In contrast, biases such as the Bandwagon Effect and Gambler's Fallacy are more sensitive to irrelevant contextual changes, as captured by elevated **AF** scores, suggesting reduced robustness against distractors.

These results demonstrate that LLM CBs are structured, not random. Salience-driven biases show dynamic, scalable distortions, outcome-related biases show adaptive evaluative shifts, and belief perseverance manifests as a mix of rigidity and contextual fragility. HR, BD, AF offer a quantitative framework to capture these mechanisms, linking LLM reasoning with human decision-making.

Relation to model architecture. Mixture-of-Experts (MoE) models, including GPT-5.4, Qwen3, and DeepSeek, generally show **positive BD** under Group B compared to dense architectures. This may reflect the selective activation of specialised subnetworks, which can inadvertently amplify context-sensitive or framing-dependent reasoning pathways. In contrast, dense models such

as Claude and Gemma tend to exhibit more rigid and stable behaviour across prompt variations, suggesting more uniform processing of perturbations.

Models using RLHF, RLAIF, or Constitutional AI (CAI) (Kaufmann et al., 2025) are generally robust to irrelevant perturbations, particularly for Group A. For example, Claude models consistently show little distractibility (**negative AF**) across both bias groups, but are highly dynamic and amplification-prone (**positive BD**) for Group B, suggesting alignment mostly improves invariance to noise rather than suppressing socially framed or retrospective heuristic effects. In contrast, less specified alignment pipelines show both higher distractibility (**positive AF**) and stronger amplification sensitivity. Coda-Forno et al. (2024) note that RLHF generally induces conservative, hyperrational behaviour, but our evaluation reveals this is context-dependent, as while RLHF protects against heuristic capture (Group A), it increases vulnerability and cognitive volatility when models are faced with narrative based, post-hoc biases (Group B). This likely arises from the optimization objectives of alignment training (Ouyang et al., 2022): objectivity guards against simple mathematical or logical traps, but the mandate to provide comprehensive explanations inadvertently forces the model to over-justify known outcomes or social frames, amplifying narrative driven reasoning errors.

Model size primarily improves AF and reduces distractibility. Larger models, particularly within the LLama and Qwen families, show stronger (negative) AF and reduced sensitivity to irrelevant information. However, scaling does not eliminate HR, as larger models remain highly sensitive to directional amplification under Group B. Thus, size improves robustness to noise more than it reduces susceptibility to meaningful bias induction.

The results show CBs in LLMs emerge from interactions between architecture, alignment, and scale instead of any single factor. Importantly, socially and retrospectively framed biases (Group B) are more resistant to mitigation than simpler heuristic perturbations (Group A), as they more effectively reshape internal reasoning trajectories rather than merely perturbing surface-level responses.

8 Conclusion

We introduced a construct-based evaluation framework for cognitive biases in LLMs, combining MT with latent variable modelling to move beyond aggregate benchmark scores. Across models and bias conditions, we find that cognitive biases are better understood as structured behavioural patterns rather than isolated errors. We identify three latent dimensions, AF, HR, and BD, which jointly describe how models respond to heuristic cues, invariance under perturbation, and amplify or stabilise biased behaviour. Results highlight the importance of evaluation at the level of behavioural constructs rather than task level.

Future work will extend this framework across a wider range of model families and architectures, expand the space of MT to include more fine-grained behavioural probes, and integrate mechanistic interpretability techniques to examine whether these latent behavioural dimensions correspond to identifiable internal activation patterns. Additionally, alternative latent variable frameworks such as Item Response Theory could be explored to model bias sensitivity at the item and prompt level.

Limitations

This study proposes a construct-based framework for evaluating cognitive biases in LLMs, however, several limitations should be acknowledged.

(1) The dataset is restricted to probabilistic tasks represented on 1-7 response scale. While this design enables standardized quantitative comparison across models and conditions, it may not fully capture the complexity of open-ended reasoning or real-world decision-making scenarios. Although the dataset is grounded in established cognitive bias literature, the constructed prompts and metamorphic transformations were not formally validated by domain experts in cognitive science or experimental psychology.

(2) The metamorphic transformations on amplitude manipulations (DIR_low and DIR_high) were constructed qualitatively by varying the semantic strength of the words used in the heuristic cues. Because these variations were not systematically calibrated or validated through human psychometric norming, the exact distance between low and high amplitude is not strictly uniform across all biases. This qualitative approach to perturbation may introduce variance into the features. Additionally, the engineered behavioural features rely on manually designed mathematical formulations. While theoretically motivated, alternative definitions of susceptibility, robustness, or elasticity could produce different latent structures.

(3) LLMs are highly sensitive to prompt phrasing. While our metamorphic testing systematically alters the heuristic cue, the underlying task templates remain fixed. Future work should evaluate whether the observed latent factors are robust across diverse syntactic templates and domains.

(4) LLMs are trained on vast corpora that include cognitive psychology literature, they may possess meta-knowledge of the biases being tested. To mitigate direct memorization, our dataset was custom-generated by designing prompts that structurally align with bias manipulation mechanics, rather than reusing classic experiments. However, there remains a risk that highly capable models recognize the underlying theoretical patterns of the prompt. This meta-awareness could trigger alignment interventions that distort the natural expression of the bias, making it difficult to separate true reasoning failure from safety-driven overcorrection.

(5) The MFT condition approximates an unbiased reference state. However, there is no guarantee that baseline responses are themselves free from hidden heuristic effects inherited from pre-training data or alignment procedures. As a result, measured deviations may underestimate or distort the true magnitude of bias behaviour.

(6) Finally, the proposed framework evaluates behavioural responses rather than internal cognitive mechanisms. Although the observed patterns functionally resemble human CBs, the results should not be interpreted as evidence that LLMs possess human-like cognition or psychological states. The analysis captures statistical and behavioural regularities in model outputs under controlled prompt transformations. The selected set of CBs, while diverse, is not exhaustive, and other forms of reasoning distortion were not included. Therefore, the latent factors identified in this work represent a functional behavioural mapping rather than a complete taxonomy of artificial cognition.

References

Ahmed Ahmed, A. Feder Cooper, Sanmi Koyejo, and Percy Liang. 2026. Extracting books from production language models. *Preprint*, arXiv:2601.02671.

Long and Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dan Hendrycks, Ziwen Han, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, and 1081 others. 2026. A benchmark of expert-level academic questions to assess ai capabilities. *Nature*, 649(8099):1139–1146.

Solomon E Asch. 1951. Effects of group pressure upon the modification and distortion of judgments. In Harold Guetzkow, editor, *Groups, leadership and men*, pages 222–236. Carnegie Press.

Jonathan Baron and John C Hershey. 1988. Outcome bias in decision evaluation. *Journal of personality and social psychology*, 54(4):569.

Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. Bad is stronger than good. *Review of general psychology*, 5(4):323–370.

Dillon Bowen III and 1 others. 2024. Measuring political bias in large language models. *arXiv preprint*.

Timothy A Brown. 2015. *Confirmatory factor analysis for applied research*. Guilford publications.

John Burden, Marko Tešić, Lorenzo Pacchiardi, and José Hernández-Orallo. 2025. Paradigms of ai evaluation: Mapping goals, methodologies and culture. *Preprint*, arXiv:2502.15620.

Ryan Burnell and 1 others. 2023. Rethinking benchmarking in nlp. *Transactions of the Association for Computational Linguistics*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. 2024. Cogbench: a large language model walks into a psychology lab. *Preprint*, arXiv:2402.18225.

Anna B Costello and Jason Osborne. 2005. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, 10(1).

Rachel Croson and James Sundali. 2005. The gambler's fallacy and the hot hand: Empirical data from casinos. *Journal of risk and uncertainty*, 30(3):195–209.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, and 1 others. 2023. Faith and fate: Limits of transformers on compositionality. *Advances in neural information processing systems*, 36:70293–70332.

Jens Echterhoff, Yuhan Liu, Abdulaziz Alessa, Julian McAuley, and Zhiting He. 2024. Cognitive bias in decision-making with llms. *Preprint*, arXiv:2403.00811.

Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, and Erin J Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272.

Baruch Fischhoff. 1975. Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human perception and performance*, 1(3):288–299.

Thomas Gilovich, Dale Griffin, and Daniel Kahneman. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press, Cambridge, UK.

Peter Gostev. 2026. Bullshitbench: A benchmark measuring whether ai models challenge nonsensical prompts. https://petergpt.github.io/bullshit-benchmark/viewer/index.v2.html. GitHub repository: https://github.com/petergpt/bullshit-benchmark.

Martie G. Haselton, Daniel Nettle, and Damian R. Murray. 2015. *The Evolution of Cognitive Bias*, chapter 41. John Wiley & Sons, Ltd.

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.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Daniel Kahneman and Amos Tversky. 1973. On the psychology of prediction. *Psychological Review*, 80(4):237–251.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2025. A survey of reinforcement learning from human feedback. *Preprint*, arXiv:2312.14925.

R.B. Kline. 2005. *Principles and Practice of Structural Equation Modeling*, 2nd edition. Guilford Press, New York.

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, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Elizabeth F Loftus and John C Palmer. 1974. Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of verbal learning and verbal behavior*, 13(5):585–589.

Sven Malberg, Roman Poletukhin, Christian M. Schuster, and Georg Groh. 2024. A comprehensive evaluation of cognitive biases in llms. *Preprint*, arXiv:2410.15413.

Richard Nadeau, Edouard Cloutier, and J-H Guay. 1993. New evidence about the existence of a bandwagon effect in the opinion formation process. *International Political Science Review*, 14(2):203–213.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Leyan Pan, Vijay Ganesh, Jacob Abernethy, Chris Esposo, and Wenke Lee. 2024. Can transformers reason logically? a study in sat solving. *arXiv preprint arXiv:2410.07432*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4902–4912.

David Rozado. 2024. The political preferences of llms. *PLOS ONE*.

Paul Rozin and Edward B Royzman. 2001. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4):296–320.

Sergio Segura, Gordon Fraser, Ana B Sanchez, and Antonio Ruiz-Cortés. 2016. A survey on metamorphic testing. *IEEE Transactions on software engineering*, 42(9):805–824.

Aman Shaikh, Rohan A. Dandekar, Suhas Panat, and Raj Dandekar. 2024. Cbeval: A framework for evaluating and interpreting cognitive biases in llms. *Preprint*, arXiv:2412.03605.

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, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Amos Tversky and Daniel Kahneman. 1971. Belief in the law of small numbers. *Psychological bulletin*, 76(2):105–110.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

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

Yixin Wan and 1 others. 2023. Jailbreak and bias analysis of large language models. *arXiv preprint*.

Peter C Wason. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12(3):129–140.

A Cognitive Bias

This appendix provides descriptions of the cognitive biases included in the evaluation and outlines how each bias was operationalized through metamorphic prompt construction. The selected biases were chosen based on their prominence in cognitive psychology and prior studies of heuristic reasoning and decision-making under uncertainty.

Anchoring Bias captures the tendency to rely on an initial reference value when making judgments, with subsequent estimates insufficiently adjusted away from that anchor. Classic experiments show that even arbitrary or irrelevant numerical anchors systematically shift estimates toward the provided value (Tversky and Kahneman, 1974). In our setting, the bias is induced by introducing an initial reference estimate within the scenario, the MFT condition contains no anchor, whereas the BIAS condition includes a suggestive starting value.

Base Rate Neglect describes the tendency to underweight prior probabilities in favour of individuating or case-specific information. In classic studies, participants are given both statistical base rates and descriptive details, yet judgments are disproportionately influenced by the latter (Kahneman and Tversky, 1973). Here, scenarios include both sources of information, with the BIAS condition emphasizing individuating evidence that may conflict with the base rate.

Bandwagon Effect reflects the influence of perceived majority opinion on individual judgment. Experimental evidence shows that individuals may adopt incorrect answers when they believe others endorse them (Asch, 1951; Nadeau et al., 1993). This effect is induced by introducing information about prevailing opinions, absent in MFT but present in BIAS, with varying strength across conditions.

Confirmation Bias involves favouring information that aligns with prior beliefs while discounting contradictory evidence (Nickerson, 1998). Classic paradigms demonstrate systematic overweighting of confirmatory information (Wason, 1960). The setup introduces an initial hypothesis in the BIAS condition, guiding interpretation relative to the neutral MFT scenario.

Framing Effect arises when equivalent information leads to different decisions depending on whether it is presented in a positive or negative form. Foundational studies show shifts between risk-averse and risk-seeking behaviour under gain versus loss frames (Tversky and Kahneman, 1981). This is implemented by presenting logically equivalent scenarios with positive (MFT) and negative (BIAS) wording.

Gambler's Fallacy captures the mistaken belief that independent random events are influenced by prior outcomes. Empirical studies show that people expect reversals after streaks of identical results (Tversky and Kahneman, 1971; Croson and Sundali, 2005). The design introduces outcome sequences in the BIAS condition, with varying streak lengths, while MFT contains no prior history.

Negativity Bias reflects the stronger influence of negative information relative to equally informative positive content. Experimental evidence shows that negative stimuli disproportionately affect judgment and memory (Baumeister et al., 2001; Rozin and Royzman, 2001). Scenarios are constructed such that BIAS emphasizes negative evidence, while MFT remains balanced.

Hindsight Bias describes the tendency to perceive outcomes as more predictable once they are known. Studies consistently show inflated estimates of prior predictability after outcome revelation (Fischhoff, 1975). This is captured by revealing the outcome in the BIAS condition while keeping it hidden in MFT.

Outcome Bias occurs when decisions are evaluated based on their results rather than the quality of the reasoning process. Experimental findings show that identical decisions are judged differently depending on outcomes (Baron and Hershey, 1988). In our setup, outcome information is introduced only in the BIAS condition.

Suggestibility reflects the incorporation of external suggestions into judgments, even when they are non-diagnostic. Research in memory and interrogation contexts shows that leading statements can systematically alter responses (Loftus and Palmer, 1974). The BIAS condition introduces such a suggestion, while MFT presents only task-relevant information.

Von Restorff effect reflects the tendency for distinctive or isolated information to receive greater

attention and influence than otherwise similar information. The von Restorff effect, also known as the isolation effect, shows that information that stands out from its surrounding context is more likely to be noticed and recalled. The design tests this effect by presenting selected information in bold in the BIAS condition, making it visually distinctive relative to the surrounding text, while MFT presents the same task without any emphasized information.

The selected cognitive biases were further organized according to the type of metamorphic manipulation used during evaluation. Group A biases employ directional strength perturbations (`DIR_high` and `DIR_low`) to examine whether biased behaviour scales with stronger or weaker heuristic cues. Group B biases use opposing perturbations (`OPP`) that reverse the framing or evaluative direction of the original bias signal while preserving the underlying task structure. Certain biases, such as the Framing Effect and Von Restorff Effect, were not naturally compatible with either perturbation strategy and were therefore evaluated separately.

Group A Biases (DIR_High / DIR_Low)
Anchoring
Bandwagon Effect
Base-rate Fallacy
Confirmation Bias
Gambler's Fallacy
Negativity Bias
Group B Biases (OPP)
Hindsight Bias
Outcome Bias
Suggestibility

Table 4: Cognitive bias groups included in the evaluation and the individual biases assigned to each group.

B Latent Factor Analysis Diagnostics

This section summarizes the statistical diagnostics used to evaluate the suitability of the engineered behavioural features for latent variable modelling.

B.1 Feature Descriptive Statistics

The engineered behavioural features exhibit substantial variability across models and bias conditions, indicating meaningful behavioural diversity suitable for latent structure analysis. Table 5 summarizes the distributional properties of each feature.

Feature	**Mean(Std)**	**Range**
Susceptibility	0.02 (0.36)	[-1, 1]
INV (Contextual)	0.36 (0.81)	[0, 6]
INV (Temporal)	0.32 (0.77)	[0, 6]
INV (Metadata)	0.32 (0.76)	[0, 6]
INV (Numerical)	0.36 (0.80)	[0, 6]
M_COG_mag	0.04 (0.98)	[-6, 6]
M_COG_dir	-0.01 (0.72)	[-1, 1]
Elasticity	2.36 (2.11)	[0, 11]
Asymmetry	0.81 (1.96)	[-6, 6]
Directional_Asym	0.95 (2.53)	[-6, 6]
Variability	0.94 (0.75)	[0, 4.24]

Table 5: Descriptive statistics of the engineered behavioural features across all tasks and evaluated model, including the mean, standard deviation, and observed value range (minimum-maximum) for each feature.

B.2 KMO and Bartlett Test Results

To assess the suitability of the feature matrices for factor analysis, we computed the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity for each evaluation group. The observed KMO values indicate acceptable factorability, while Bartlett's test confirms that the correlation matrices significantly differ from identity matrices (Table 6).

Group	**KMO**	**Bartlett's Test (χ^2)**
Group A	0.698	24464.44
Group B	0.623	15287.35
All	0.731	23900.64

*** $p < 0.001$

Table 6: Kaiser-Meyer-Olkin (KMO) sampling adequacy and Bartlett's test of sphericity for each feature group.

B.3 Factor Selection

Factors were retained when observed eigenvalues exceeded both the Kaiser criterion (> 1) and the corresponding random eigenvalues estimated through parallel analysis (see Figure 5).

C Prompt Template

This section presents examples of the prompt templates used to evaluate the language models. It first describes the types of invariance incorporated into

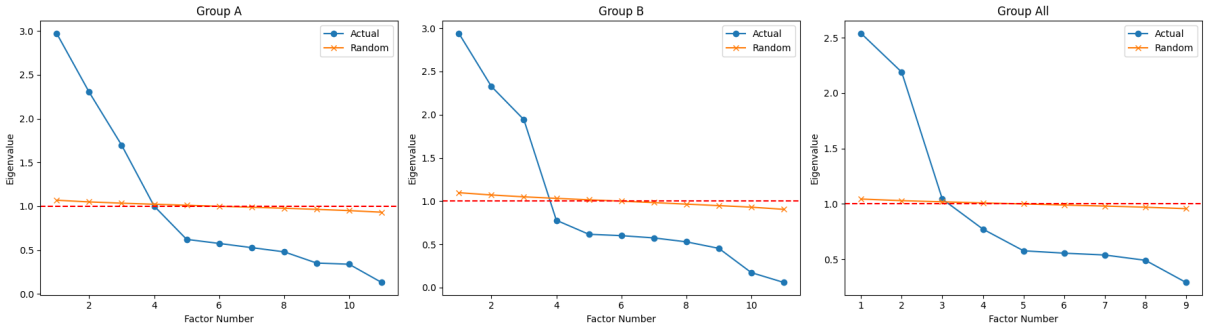


Figure 5: Scree plots with parallel analysis for Group A (DIR), Group B (OPP), and the shared feature set for all bias, illustrating latent factor selection based on observed and random eigenvalues.

the evaluation design, together with representative examples illustrating how these invariances are applied across prompts. Then full prompt examples for biases from Group A and Group B, showing how the corresponding bias conditions are constructed. These examples illustrate the structure of the evaluation prompts and the controlled differences between conditions while keeping the underlying task and relevant information consistent.

C.1 Types of invariance test

Invariance tests are used to assess the robustness of heuristic reasoning to information that is irrelevant to the underlying task. Specifically, they examine whether model responses change when non-essential details are modified while the core task, relevant evidence, and expected reasoning remain unchanged. A robust model should produce consistent responses under such perturbations, indicating that its reasoning is not unduly influenced by incidental contextual information or noise.

We consider four types of invariance tests, each targeting a different form of irrelevant variation: *INV*, which modifies general contextual details; *INV_TEMP*, which varies temporal information; *INV_META*, which changes metadata or document-related information; and *INV_NUM*, which alters numerical details that are not relevant to the task. Table 7 provides representative examples of each invariance type.

C.2 Group A bias Example: Confirmation Bias

The perturbations were designed to preserve the underlying reasoning structure and normative solution while introducing semantically irrelevant modifications. For each task instance, perturbations were inserted as localized modifications to the original prompt. Multiple template variants were used for each perturbation type to increase contextual diversity while preserving the semantic irrelevance of the modifications to the target reasoning task. Representative examples of these invariance perturbations and their corresponding prompt templates are provided in Table 8.

INV Type	Example
INV	The patient was seated in a room with beige walls. The patient was seated in a room with light blue walls.
INV_TEMP	The case was reviewed on a Monday. The case was reviewed on a Wednesday.
INV_META	The document was transcribed from a voice recording. The record was stored in a cloud-based system.
INV_NUM	The report is four pages long. The review took place in a building with 10 floors.

Table 7: Representative invariance perturbations used in the evaluation framework.

C.3 Group B bias Example: Suggestibility Bias

We implemented four classes of invariance perturbations: contextual (`INV`), temporal (`INV_TEMP`), metadata (`INV_META`), and numerical (`INV_NUM`). These perturbations were designed to preserve the underlying reasoning structure and normative solution while introducing semantically irrelevant modifications. For each task instance, perturbations were inserted as localized modifications to the original prompt. Representative examples of these invariance perturbations and their corresponding prompt templates are provided in Table 9.



Figure 6: Latent behavioural representation of cognitive biases across evaluation groups. The left panel shows Group A biases evaluated using directional strength perturbations (`DIR_high`/`DIR_low`), while the right panel shows Group B biases evaluated using opposing perturbations (`OPP`).

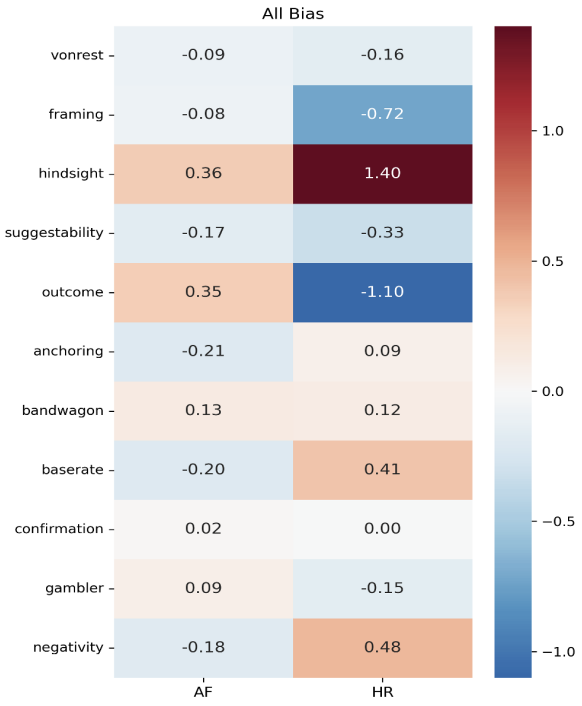


Figure 7: Latent behavioural mapping of cognitive biases

D Bias Factor Score

This section presents the factor scores associated with the evaluated cognitive biases, illustrating their latent behavioural profiles across the extracted factors. We identify three latent dimensions, AF, HR, and BD, which jointly describe how models respond to heuristic cues, invariance under perturbation, and amplify or stabilise biased behaviour. Figure 7 shows the factor scores for all evaluated biases, while Figure 6 presents the corresponding factor scores separately for Group A and Group B biases. These visualizations provide an overview of how the evaluated biases are positioned across the underlying behavioural dimensions.

E Model Factor Scores

Figures 8 and 9 present the factor scores for individual models across model families and evaluation groups. Each subplot corresponds to a model family, with the three rows denoting the AF, HR, and BD evaluation dimensions. The horizontal axis represents the factor score, and the dashed vertical line marks the zero reference point. Different colours distinguish model variants within each family, enabling comparison of the direction and magnitude of factor scores across models. The plots provide a more granular view of model-level variation.

F Structural equation modelling

To validate the latent behavioural structure identified during exploratory factor analysis, we conducted confirmatory factor analysis (CFA) and structural equation modelling (SEM) for the Group B feature set. The hypothesized model consisted of three latent constructs: Attentional Filtering (AF), Heuristic Reliance (HR), and Bias Dynamics (BD), with observed indicators assigned according to the exploratory factor structure. Model fit was evaluated using standard SEM goodness-of-fit indices.

Table 10 summarizes the resulting fit statistics. Overall, the model demonstrates strong structural

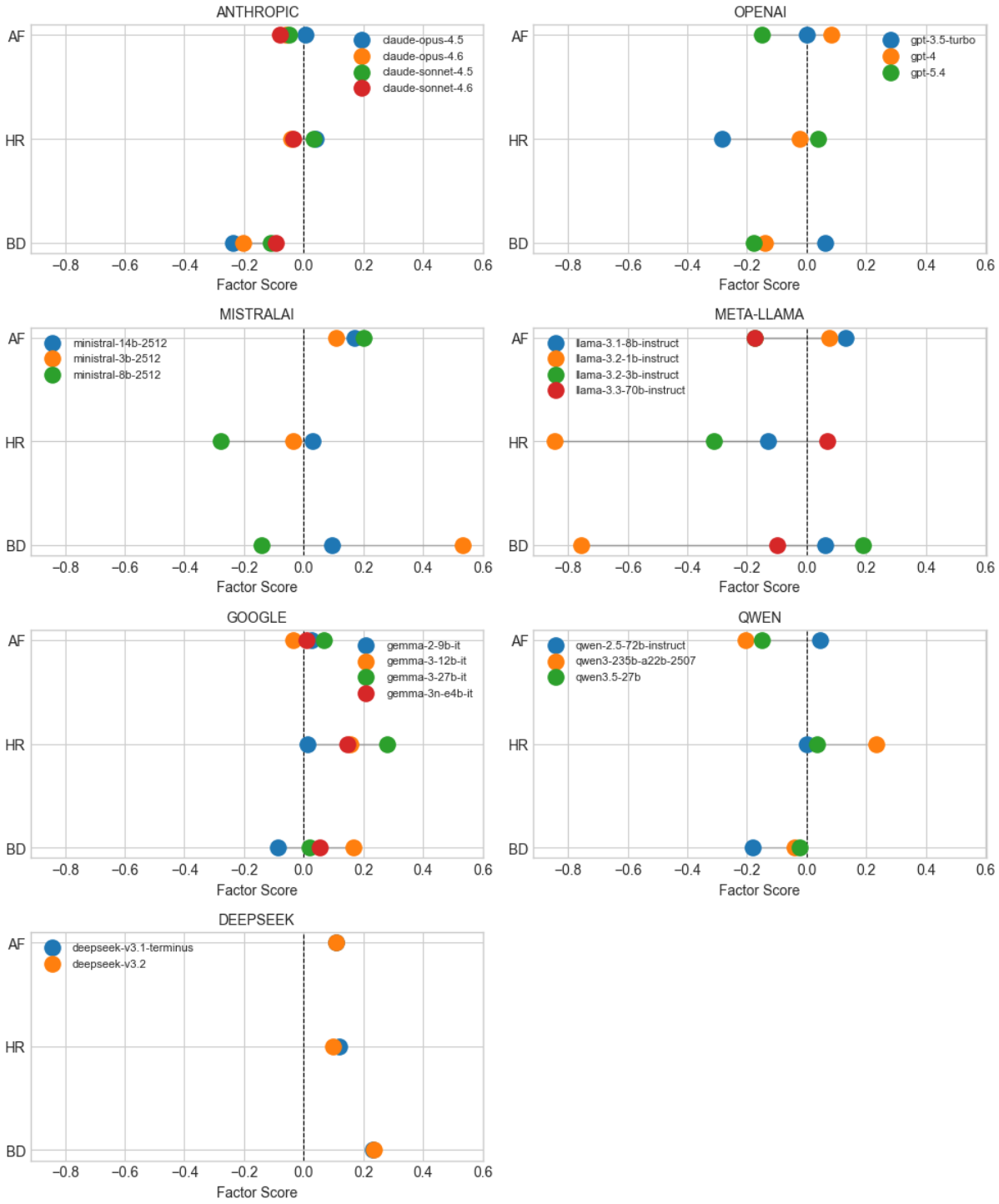


Figure 8: Model-level latent factor scores for Group A across model families and evaluation dimensions (AF, HR, and BD). The dashed line denotes a factor score of zero.

validity, with all major fit indices exceeding commonly accepted thresholds and RMSEA indicating good approximate fit.

Figure 10 illustrates the resulting structural model, including latent constructs, observed indicators, and standardized path coefficients between behavioural dimensions.

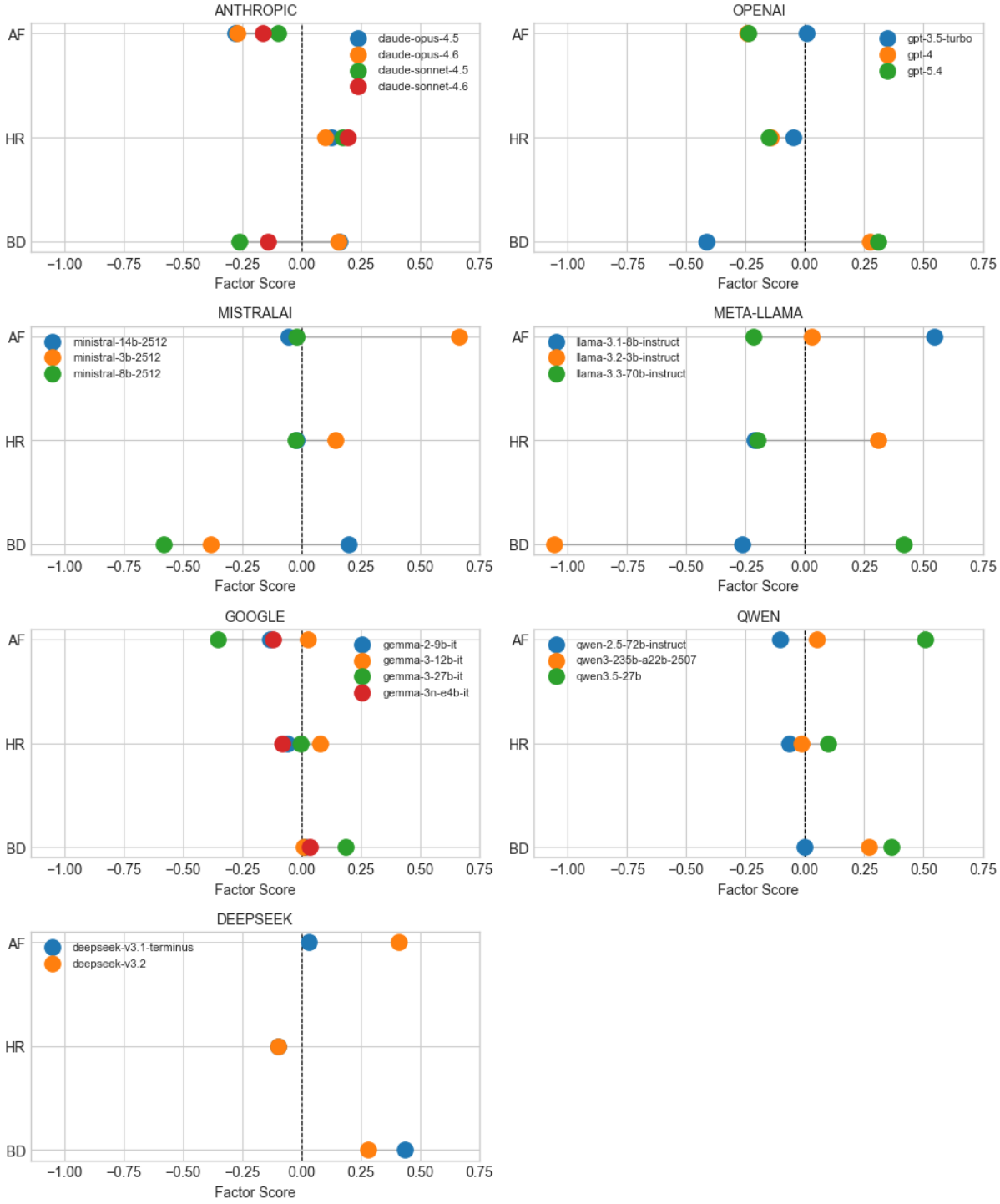


Figure 9: Model-level latent factor scores for Group B across model families and evaluation dimensions (AF, HR, and BD). The dashed line denotes a factor score of zero.

Condition	Prompt
Role	Suppose you are a medical decision maker. Your task is to evaluate the provided information and select the most appropriate option.
MFT	*A 45-year-old patient presents with fever, productive cough, and shortness of breath for 2 days.*
BIAS	*An initial clinical impression suggests this may be bacterial pneumonia. A 45-year-old patient presents with fever, productive cough, and shortness of breath for 2 days.*
INV	*An initial clinical impression suggests this may be bacterial pneumonia. A 45-year-old patient presents with fever, productive cough, and shortness of breath for 2 days. The patient was seated in a room with white walls.*
INV_TEMP	*An initial clinical impression suggests this may be bacterial pneumonia. A 45-year-old patient presents with fever, productive cough, and shortness of breath for 2 days. The case was reviewed on a Friday.*
INV_META	*An initial clinical impression suggests this may be bacterial pneumonia. A 45-year-old patient presents with fever, productive cough, and shortness of breath for 2 days. The record was stored in a cloud-based system.*
INV_NUM	*An initial clinical impression suggests this may be bacterial pneumonia. A 45-year-old patient presents with fever, productive cough, and shortness of breath for 2 days. The review took place in a building with 10 floors.*
DIR_High	*A strong clinical impression suggests this is likely bacterial pneumonia. A 45-year-old patient presents with fever, productive cough, and shortness of breath for 2 days.*
DIR_Low	*A preliminary clinical impression suggests this might be bacterial pneumonia. A 45-year-old patient presents with fever, productive cough, and shortness of breath for 2 days.*
Meta_Cog	*An initial clinical impression suggests this may be bacterial pneumonia. A 45-year-old patient presents with fever, productive cough, and shortness of breath for 2 days. Please be aware that the following case may contain a cognitive bias.*
Question	What is the probability that the patient has bacterial pneumonia?
Option	Choose one option. Respond with the option number ONLY. 1. 0–14%; 2. 15–28%; 3. 29–42%; 4. 43–57%; 5. 58–71%; 6. 72–85%; 7. 86–100%

Table 8: Full prompt variants for a medical task in the evaluation of Confirmation Bias. The MFT condition presents the clinical information without a prior directional impression, while the BIAS condition introduces a heuristic cue. The invariance conditions introduce irrelevant contextual variations while preserving the underlying task and reasoning structure. The DIR tests the the capability to increase or decrease the bias effect. The Meta_cog tests whether the model have self correction property.

Condition	Prompt
Role	*Suppose you are a medical decision maker. Your task is to evaluate the provided information and select the most appropriate option.*
MFT	*A 59-year-old treated for anaemia shows stable haemoglobin levels and improving energy. However, mild fatigue persists and appetite remains modest.*
BIAS	*A 59-year-old treated for anaemia shows stable haemoglobin levels and improving energy. However, mild fatigue persists and appetite remains modest. Given these findings suggesting the anaemia is likely stable,*
OPP	*A 59-year-old treated for anaemia shows stable haemoglobin levels and improving energy. However, mild fatigue persists and appetite remains modest. Given these findings suggesting the anaemia may be worsening,*
INV	*A 59-year-old treated for anaemia shows stable haemoglobin levels and improving energy. However, mild fatigue persists and appetite remains modest. The patient was seated in a room with white walls. Given these findings suggesting the anaemia is likely stable,*
INV_TEMP	*A 59-year-old treated for anaemia shows stable haemoglobin levels and improving energy. However, mild fatigue persists and appetite remains modest. The case was reviewed on a Friday. Given these findings suggesting the anaemia is likely stable,*
INV_META	*A 59-year-old treated for anaemia shows stable haemoglobin levels and improving energy. However, mild fatigue persists and appetite remains modest. The record was stored in a cloud-based system. Given these findings suggesting the anaemia is likely stable,*
INV_NUM	*A 59-year-old treated for anaemia shows stable haemoglobin levels and improving energy. However, mild fatigue persists and appetite remains modest. The review took place in a building with 10 floors. Given these findings suggesting the anaemia is likely stable,*
Meta_Cog	*A 59-year-old treated for anaemia shows stable haemoglobin levels and improving energy. However, mild fatigue persists and appetite remains modest. Please be aware that the following case may contain a cognitive bias. Given these findings suggesting the anaemia is likely stable,*
Question	*What is the probability that the anaemia is worsening?*
Option	Choose one option. Respond with the option number ONLY. 1. 0–14%; 2. 15–28%; 3. 29–42%; 4. 43–57%; 5. 58–71%; 6. 72–85%; 7. 86–100%

Table 9: Full prompt variants for a medical task in the evaluation of Suggestibility. The MFT condition presents the clinical information without a prior directional impression, while the BIAS condition introduces a heuristic cue.

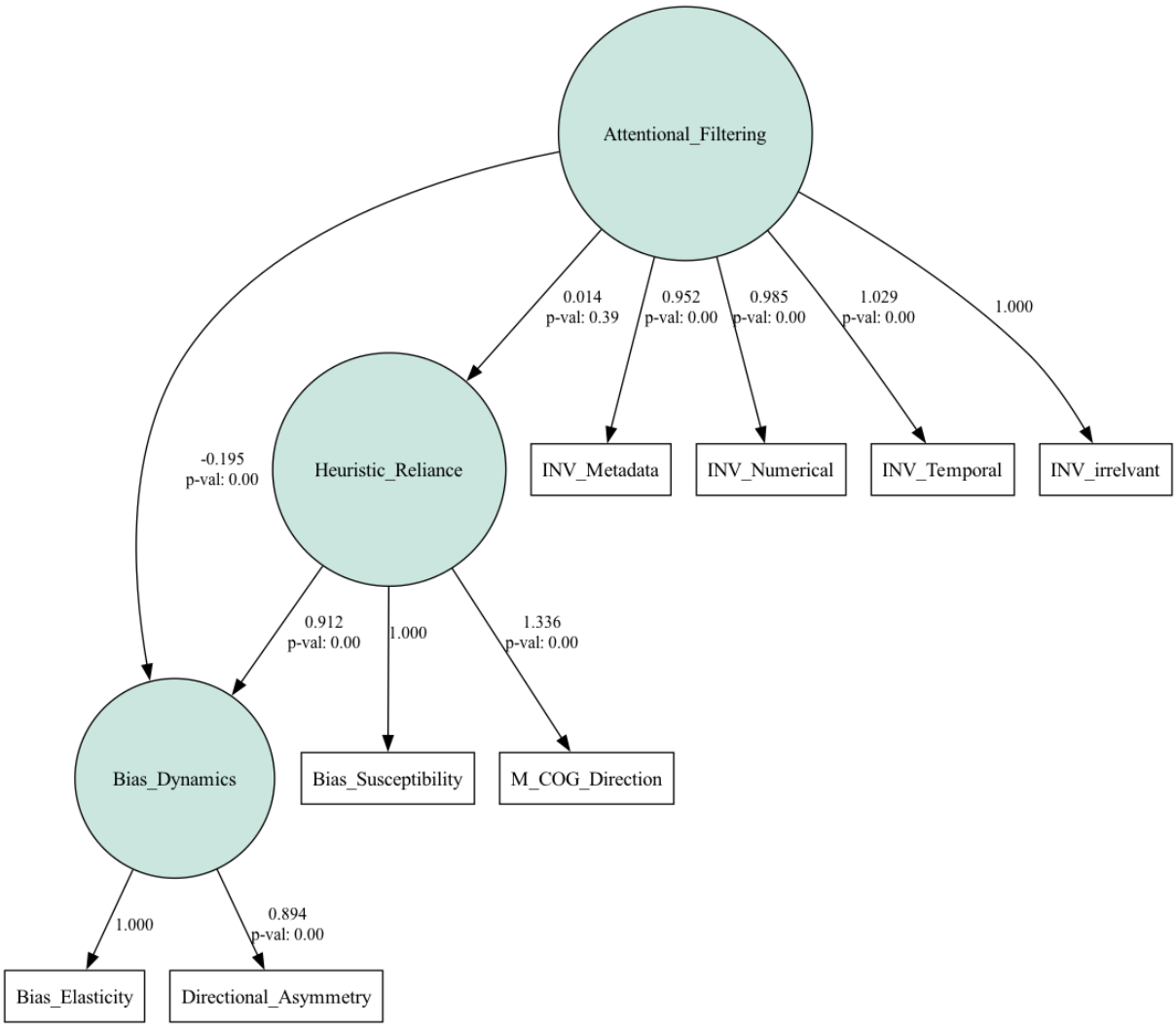


Figure 10: Structural equation model for Group B biases showing the relationships between the latent behavioural constructs. Observed indicators correspond to engineered behavioural features, while path coefficients denote standardized relationships between latent variables and indicators.

Fit Index value
χ^2 Significance = $p < 0.001$
Comparative Fit Index (CFI) = 0.994
Goodness-of-Fit Index (GFI) = 0.991
Adjusted Goodness-of-Fit Index (AGFI) = 0.985
Normed Fit Index (NFI) = 0.991
Tucker–Lewis Index (TLI) = 0.990
RMSEA = 0.027

Table 10: Goodness-of-fit statistics for the confirmatory factor analysis and structural equation model applied to the Group B behavioural feature set.