CogniBias: A Benchmark for Cognitive Biases in AI–Human Dialogue

Anonymous Author(s)
Affiliation
Address
email

Abstract

Cognitive biases are predictable departures from rational judgment that impact people's decisions and communication. As large language models (LLMs) are increasingly interfaced with everyday interactions, it is valuable to understand how biases arise and spread through AI–human conversations. We present CogniBias, the first benchmark dataset to study and assess cognitive biases in conversational settings. CogniBias is composed of 30+ established types of biases, like anchoring, framing, confirmation bias, and optimism bias, distilled from a variety of authentic question and answer scenarios. Each dialogue sample includes an LLM suggestion, a human-like response, and expert-informed annotations with notes, bias labels, and confidence scores. To establish the benchmark, we describe a generation pipeline incorporating multiple LLMs and include baseline results with pretrained and fine-tuned models on classification and detection tasks. Our analysis highlights the identified challenges in detecting subtle biases or overlapping biases and identified each model's frequent failures. We hope that by releasing CogniBias, it will align divergent perspectives on cognitive bias assessment and establish a baseline dataset for fairer, more trustworthy conversational AI systems.

1 Introduction

Cognitive biases are systematic deviations from rational judgment that shape perception, decision-making, and interaction, including well-studied cases such as anchoring, confirmation bias, framing effects, and overconfidence (6). While these insights from psychology are valuable for understanding human reasoning, they also pose risks in AI–human interaction, where conversational systems may amplify existing user biases or introduce new ones. Although recent NLP research has advanced fairness, bias detection, and debiasing, most benchmarks (e.g., WinoBias, StereoSet, CrowS-Pairs) focus on demographic or stereotypical biases (10), overlooking cognitive biases that emerge in decision-making and dialogue. Prior datasets on framing or persuasion capture isolated phenomena within narrow domains, but they lack the breadth needed to generalize across multiple categories of cognitive bias.

To address this gap, we present CogniBias, a benchmark dataset for evaluating cognitive biases in AI–human dialogue. CogniBias integrates over 30 bias types from established psychological taxonomies (e.g., anchoring, availability heuristic, sunk cost fallacy), instantiated in decision-making dialogues. Each dialogue consists of a decision-oriented question, an AI-generated suggestion, and a human-like response. Expert annotations provide bias labels and confidence scores, enabling systematic study of how biases emerge and interact in machine-mediated dialogue (8).

Our contributions are threefold:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Figure 1: Motivation for CogniBias. Conversational AI may introduce or amplify cognitive biases in user decision-making. CogniBias provides the first benchmark to systematically study such effects.

- CogniBias dataset: A large-scale, multi-bias benchmark covering 30+ types of cognitive biases in AI–human dialogue.
- Generation pipeline: A reproducible LLM-based framework combining dialogue simulation with expert-informed annotations.

By releasing CogniBias, we aim to foster cross-disciplinary research at the intersection of cognitive psychology, NLP, and human–AI interaction. We envision this resource as a foundation for building conversational AI systems that are accurate, fair, and free from cognitive distortions that undermine trust and decision quality.

2 Related work

The NLP field has developed various benchmarks which have aimed to understand and measure bias in language models. This work has, at least to date, focused on representational-type harms related to bias based on gender, race, and/or culture. For instance, StereoSet, CrowS-Pairs and WinoBias (7; 13; 10) showed that pretrained models reinforced systemic stereotypes, and inspired new research in debiasing. However, these benchmarks and the corpus based on them, only assess demographic and social stereotypes and do not measure the cognitive processes underlying reasoning and/or decision-making.

Cognitive psychology has long documented systematic deviations in judgment, such as anchoring, framing, and the availability heuristic (5), yet these cognitive biases are largely absent from computational evaluation benchmarks. Existing NLP studies on persuasion, framing, or fallacies typically focus on one or two bias types within narrow domains like news, politics, or social media, limiting their generalizability across bias taxonomies and failing to capture conversational dynamics.

Recent research (11; 12) has indicated that large language models (LLMs) not only inherit social stereotypes, but can also exacerbate cognitive biases in dialog. For example, an anchoring effect can occur when users overweight an initial suggestion from a model, and a framing effect can bias user judgment depending on which way the information is presented. Regardless of these important considerations, there is not yet a systematic benchmark on measuring cognitive biases in a human interactive AI setting. There has been a focus on ethical risk or persuasive misuse of interacting with an LLM but no empirical datasets to use for a thorough examination of those concepts.

CogniBias differs from prior bias benchmarks by addressing cognitive rather than demographic biases within dialogue contexts 1.

3 Dataset design

3.1 Bias Taxonomy

CogniBias is founded on a classification of 30+ cognitive biases situated in cognitive psychology. These include classic heuristics and judgment errors such as anchoring, confirmation bias, framing, availability heuristic, gambler's fallacy, sunk cost fallacy, halo effect, overconfidence, optimism bias, and spotlight effect, as well as various other biases. The classification was an aggregation from the psychology literature and was modified slightly for conversational contexts to ensure that biases

Table 1: Comparative analysis of bias datasets. CogniBias uniquely focuses on cognitive biases in AI–human dialogue, unlike prior benchmarks addressing demographic stereotypes.

Aspect	CogniBias	StereoSet	CrowS-Pairs	WinoBias
Primary Focus	Cognitive biases in AI–human dialogues (e.g., anchoring, framing, optimism bias)	Stereotypical social biases (gender, race, religion, profession)	Social biases across demographic categories	Gender bias in coreference resolution
Bias Type	30+ **cognitive** bias types from psychology	4 **social** bias domains	9 **social** bias categories	Gender stereotypes only
Data Format	Three-turn dialogues: *decision question* $\rightarrow$ *AI suggestion* $\rightarrow$ *human-like response*	Triplets: *stereotype / anti-stereotype / unrelated*	Paired sentences: *stereotype vs. anti-stereotype*	Sentence templates with gendered pronouns
Scale	3,000 dialogues; 30+ bias types; 50 simulated participants	16,995 instances; 321 target terms	1,508 pairs; 9 bias types	396 examples (2 test sets)
Domains Covered	Finance, health, lifestyle, and social decision-making	Gender, race, religion, profession	Race, gender, religion, age, etc.	Occupation–gender stereotypes
Interaction Type	Conversational — includes both AI and human-like turns	Static sentence or discourse	Static sentence pairs	Static coreference sentences
Uniqueness	First dataset to study cognitive biases in AI–human dialogue	Measures demographic stereotypes only	Broad demographic coverage but not conversational	Template-based; focused narrowly on gender
Human–AI Interaction	Present — AI output may influence human bias	Absent	Absent	Absent

covered a wide range of reasoning errors. We code each dialogue with exactly one main bias type (or No Bias, when appropriate) (6).

3.2 Dialogue Format

Each dataset instance follows a three-part dialogue structure:

- Question of action: A plausible situation that calls for decision-making (e.g., "Is it an opportune moment for me to invest in cryptocurrency?").
- AI proposal: The model-generated response emulates the kinds of advice, or context framing, that an LLM might use.
- Response of a human-like persona: A responding utterance replicates how a person would respond directly in a moment of cognitive bias.

Annotations include:

- Bias label (one of the 30+ categories)
- Confidence score (0–100), reflecting annotator certainty,
- Optional notes, used during quality control.

This design captures interactive dynamics where AI output may influence or reinforce human cognitive distortions.

3.3 Generation Pipeline

To create CogniBias, we implemented a two-stage generation pipeline:

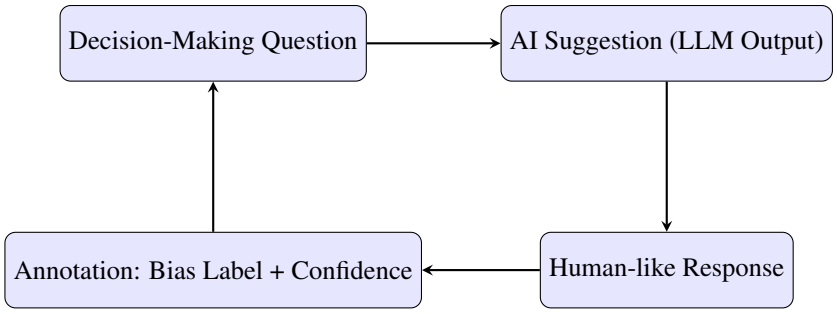

Figure 2: Overview of the CogniBias dataset generation pipeline. Each instance includes a decision-making question, an AI-generated suggestion, and a human-like response annotated with bias type and confidence score.

- Prompted LLM generation: A number of large language models (Gemini, Groq's LLaMA variants, etc.) were prompted with a pool of 60+ fallback questions covering the domains of finance, health, lifestyle, and social decision-making. The models generated both AI-generated suggestions and participant-like responses that varied stylistically and with chain of reasoning.
- Expert-informed annotation: Sampled each dialogue for one or more candidate bias from the taxonomy. Disagreement was resolved by discussion and there are now a bias label and confidence score in the final version of the annotations.

This hybrid pipeline has the benefits of leveraging the scale of LLM simulation, while using experts to maintain fidelity to psychological definitions.

3.4 Dataset Statistics

The released dataset includes:

- Participants: 50 simulated participants, each with a dialogue history (e.g., P01, P26)
- Dialogues: 3000 annotated dialogue samples across diverse domains.
- Bias coverage: All 30+ bias categories are represented, with frequency distributions available in Appendix A.
- Annotations: Every instance contains a bias type and confidence score (mean confidence 73 across the dataset).

3.5 Quality Control

Each annotation in CogniBias was validated against cognitive psychology definitions to ensure accurate bias categorization. To reduce overfitting, the dataset also includes instances labeled as No Bias. Furthermore, the questions span diverse domains such as finance, health, relationships, consumer choices, and ethics, ensuring broad applicability.

4 Challenges & Future Use

4.1 Challenges and Future Use

We highlight several open challenges for future research. Many biases are expressed subtly and implicitly, making them difficult to identify and model. Categories can overlap, as a single utterance may fit multiple bias types, and some biases occur more frequently than others, creating class imbalance. Additionally, context sensitivity complicates modeling, since the same surface form may signal different biases depending on the conversational setting. Future work could explore bias-adaptive prompting strategies, where large language models dynamically adjust their communication style to mitigate bias triggers by reframing questions, providing counter-examples, or requesting

supporting evidence. Another promising direction is to treat cognitive bias detection as a multi-task, hierarchical NLP problem that captures both sentence-level and dialogue-level phenomena, accounting for overlapping biases and leveraging relationships among cognitive bias categories. By surfacing these challenges and directions, CogniBias provides not only a benchmark but also a foundation for advancing methods in bias detection, dialogue modeling, and cognitively aware AI evaluation.

5 Discussion & Limitations

CogniBias provides a new benchmark for evaluating cognitive biases in conversations, yet several limitations should be noted. Annotation ambiguity arises since biases often overlap (e.g., framing and optimism), and we chose to label only the dominant bias to simplify the task, which sacrifices some realism in representing the complexity of human reasoning (9). Our taxonomy covers over 30 biases but does not capture the full spectrum of distortions (e.g., moral licensing, hot-cold empathy gap) due to annotation complexity and limited examples (1). Furthermore, all dialogues are LLM-generated, which, while enabling scalability and control, reflect model-specific linguistic tendencies rather than fully naturalistic human behavior (12). The dataset also has uneven domain balance, with certain topics such as consumer decisions being overrepresented compared to others like health or finance. Annotation confidence averaged around 70, highlighting both the difficulty and subjectivity of identifying cognitive biases, which in turn increases the challenge of evaluation. Despite these limitations, CogniBias remains a systematic and scalable resource to study bias in dialogue, offering a foundation for research on bias detection, conversational fairness, and evaluation of language models grounded in psychological theory.

6 Broader Impact & Ethics

CogniBias is released to support research on how cognitive biases emerge in AI–human dialogue, with the aim of improving detection tools, bias-aware training, and the trustworthiness of conversational systems (11). While such a resource can advance transparency, equity, and reliability, it also carries risks of misuse, such as exploiting anchoring or framing for manipulation. To mitigate this, the dataset is framed strictly for evaluation and responsible use (12). Since the dialogues are simulated with LLMs and expert annotation, they do not fully capture cultural or contextual diversity, but future work could expand coverage across languages and settings (2). Importantly, no personal data are included, avoiding privacy concerns associated with scraped corpora (4). With these safeguards, CogniBias is intended to encourage responsible progress in conversational AI research while acknowledging both its potential and limitations. By situating biases in dialogue explicitly, it also invites interdisciplinary collaboration between AI, psychology, and ethics. This, in turn, can help inform design guidelines for safer systems. Ultimately, we hope CogniBias will serve not just as a benchmark but as a foundation for building more equitable and trustworthy AI dialogue technologies.

7 Conclusion

We present CogniBias, the first benchmark dataset designed for studying cognitive biases in AI–human dialogues. It includes over 30 types of biases identified in psychology and about 3,000 annotated dialogues, each consisting of a decision-oriented question, an AI recommendation, and a human-like response labeled for bias type and confidence. CogniBias enables systematic investigation of how biases manifest in conversational contexts and supports the development of bias detection or mitigation models. Our analysis highlights challenges such as overlapping categories, annotation ambiguity, and the complexity of conversational cues. By profiling cognitive biases in a structured way, CogniBias aims to inform future research across cognitive psychology, natural language processing, and human–AI interaction, fostering the development of conversational AI systems that are not only fluent and accurate but also fair, trustworthy, and aligned with human reasoning.

References

[1] Haselton, M. G., Nettle, D., & Andrews, P. W. (2015). The evolution of cognitive bias. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 724–746).

[2] R. Segerer. Cultural value alignment in large language models: A prompt-based analysis of Schwartz values in Gemini, ChatGPT, and DeepSeek. *arXiv preprint arXiv:2505.17112*, 2025.

[3] Singh, S., Singla, Y. K., Si, H., & Krishnamurthy, B. (2024). Measuring and improving persuasiveness of large language models. *arXiv preprint arXiv:2410.02653*.

[4] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 1–10.

[5] Kahneman, D. (2011). *Fast and slow thinking*. Allen Lane and Penguin Books, New York.

[6] Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.

[7] Asad, N., Sahoo, N. R., Murthy, R., Nath, S., & Bhattacharyya, P. (2025, July). "You are Beautiful, Body Image Stereotypes are Ugly!" BIStereo: A Benchmark to Measure Body Image Stereotypes in Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 24471–24496).

[8] Chen, H., Ghosal, D., Majumder, N., Hussain, A., & Poria, S. (2021). Persuasive dialogue understanding: The baselines and negative results. *Neurocomputing*, 431, 47–56.

[9] Pavlick, E., & Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7, 677–694.

[10] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and mitigation. *EMNLP*, 2018.

[11] Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The Woman Worked as a Babysitter: On Bias in Language Models. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3407–3412.

[12] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623.

[13] Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

A Bias Taxonomy

Table 2 lists the cognitive biases covered in COGNIBIAS with concise definitions.

Bias Type	Definition
Anchoring	Relying too much on the first piece of information encountered.
Availability Heuristic	Judging likelihood by easily recalled examples.
Confirmation Bias	Seeking or interpreting information that confirms prior beliefs.
Framing Effect	Making different decisions depending on how information is presented.
Hindsight Bias	Viewing past events as more predictable than they were.
Loss Aversion	Preferring to avoid losses rather than acquire equivalent gains.
Status Quo Bias	Preferring things to stay the same instead of changing.
Optimism Bias	Overestimating the likelihood of positive outcomes.
Pessimism Bias	Overestimating the likelihood of negative outcomes.
Bandwagon Effect	Adopting beliefs or behaviors because others do.
Sunk Cost Fallacy	Continuing due to prior investments despite poor prospects.
Gambler's Fallacy	Believing past random events affect future probabilities.
Overconfidence	Overestimating one's knowledge, ability, or predictions.
Halo Effect	Letting one positive trait influence overall judgment.
Self-Serving Bias	Attributing success to oneself, failures to external factors.
Dunning–Kruger Effect	Overestimating competence due to limited knowledge.
Negativity Bias	Giving greater weight to negative experiences over positive ones.
Survivorship Bias	Focusing on successful cases while ignoring failures.
Authority Bias	Placing undue weight on the opinions of authority figures.
Recency Bias	Overemphasizing recent events compared to older information.
Outcome Bias	Judging decisions by their outcomes instead of reasoning quality.
Planning Fallacy	Underestimating the time or resources needed for tasks.
Spotlight Effect	Overestimating how much others notice one's actions or appearance.
Illusory Correlation	Perceiving a relationship between unrelated events.
Base Rate Fallacy	Ignoring statistical base rates in favor of anecdotal evidence.

Table 2: Concise taxonomy of cognitive biases in COGNIBIAS.

B Dataset Statistics

Table 3 reports the frequency and average annotator confidence for all 30+ cognitive biases in COGNIBIAS.

B.1 Domain Distribution

The dataset currently covers four domains: finance, health, lifestyle, and social decision-making. In this release, domain labels are under-specified (annotated as "unknown" in the JSON files), but prompts were designed to ensure cross-domain diversity. Future expansions will include explicit domain annotations for finer-grained analysis.

C Annotation Guidelines

This section summarizes the annotation instructions provided to expert annotators for labeling dialogues in COGNIBIAS. The goal was to ensure consistent application of cognitive bias categories across 30+ types drawn from psychology.

C.1 General Instructions

Annotators were asked to:

- Read the dialogue in full (decision-oriented question, AI suggestion, and human-like response).
- Identify whether the human-like response exhibits a cognitive bias.

Bias Type	Count	Avg. Confidence
Pessimism Bias	121	60.4
Optimism Bias	110	82.4
Framing Effect	108	74.8
Spotlight Effect	108	66.0
No Bias	104	70.8
Hindsight Bias	103	64.4
Recency Bias	100	65.0
Planning Fallacy	98	53.8
Illusion of Control	96	65.3
Status Quo Bias	95	67.0
Confirmation Bias	95	87.2
False Consensus Effect	95	63.9
Selection Bias	94	64.7
Pro-Innovation Bias	93	65.6
Outcome Bias	92	65.6
Just-World Hypothesis	91	65.0
Sunk Cost Fallacy	91	65.9
Dunning–Kruger Effect	90	94.2
Gambler's Fallacy	89	64.0
Halo Effect	87	77.1
Loss Aversion	86	79.0
Bandwagon Effect	85	74.9
Negativity Bias	84	64.3
Authority Bias	84	82.0
Actor–Observer Bias	82	67.0
Self-Serving Bias	81	87.0
Base Rate Fallacy	81	63.8
Overconfidence	80	92.0
Illusory Correlation	80	65.4
Anchoring	79	77.7
Availability Heuristic	78	69.7
Group Attribution Error	71	65.5
Survivorship Bias	69	65.0

Table 3: Frequency and average annotator confidence for all cognitive bias categories in CogniBias.

- If biased, assign exactly one **dominant bias type** from the taxonomy (see Appendix A).
- If the response was neutral, label it as **No Bias**.
- Assign a **confidence score** (0–100) indicating certainty in the label.
- Provide optional notes explaining the rationale in ambiguous cases.

C.2 Bias Labeling Criteria

- **Dominant bias rule:** When multiple biases could apply, annotators selected the one most central to the response.
- **Surface vs. latent cues:** Explicit linguistic markers (e.g., "since it was first mentioned") were prioritized, but pragmatic and discourse-level reasoning (e.g., optimism bias without explicit markers) was also considered valid.
- **No Bias category:** Responses that were rational, balanced, or contextually neutral were marked as *No Bias*.

C.3 Resolving Disagreements

- Each dialogue was annotated independently by two experts.
- Disagreements were discussed in weekly calibration meetings.
- In cases of persistent disagreement, the annotation was adjudicated by a third expert.
- Confidence scores were averaged across annotators for the final dataset.

These guidelines ensured that annotations were consistent, reproducible, and grounded in established cognitive psychology definitions while acknowledging the inherent subjectivity in bias identification.

D Dataset Format

The COGNIBIAS dataset is released under an open license for research purposes. It is provided in JSON format with one file per participant. Each file contains a list of dialogues, where each dialogue is annotated with bias type and confidence.

D.1 Fields

Each dialogue instance includes the following fields:

- `participant_id` – Unique identifier for the participant (e.g., `P01`).
- `question_id` – Identifier for the dialogue question.
- `question` – Decision-making prompt shown to the participant.
- `ai_suggestion` – Response generated by a language model.
- `human_response` – Human-like response reflecting possible bias.
- `bias_type` – Assigned label from the taxonomy (Appendix A).
- `confidence` – Annotator confidence score (0–100).

D.2 Example JSON Snippet

Example JSON Snippet

```
{
  "participant_id": "P01",
  "dialogue": [
    {
      "question_id": "Q15",
      "question": "Should I invest in cryptocurrency right now?",
      "ai_suggestion": "It could be profitable, but it is highly
          volatile.",
      "human_response": "Yes,  I ve  heard so many success stories
          recently. I  d o n t  want to miss out.",
      "bias_type": "Availability Heuristic",
      "confidence": 78
    }
  ]
}
```

E Distribution of Bias Types

CogniBias encapsulates over 30 unique cognitive biases, amounting to 3,000 annotated dialogues collected from 50 participants. Participants showed an average confidence of 70.6 on a scale of 0–100 for each annotation. The most common biases included: Pessimism Bias, Optimism Bias, Spotlight Effect, Framing Effect, and No Bias (all present in >100 dialogues). Important, but less frequent biases included: Halo Effect, Planning Fallacy, Authority Bias, and Illusory Correlation. Figure 3 displays the frequency of the top 15 most common bias types.

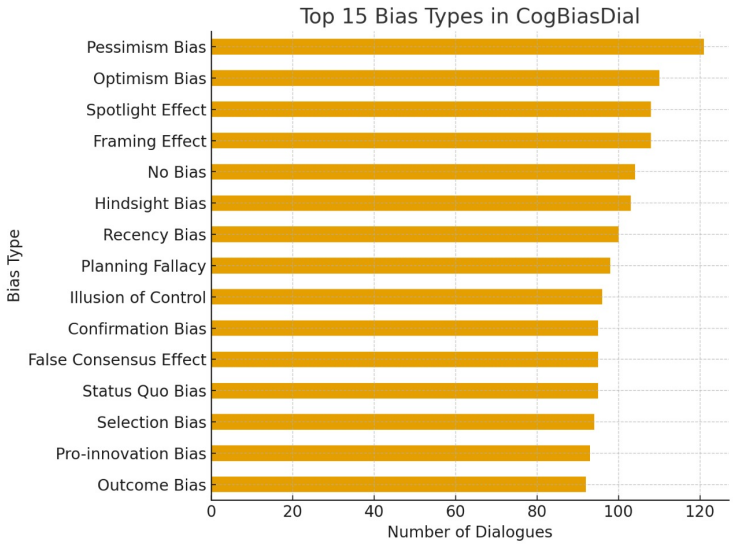

Figure 3: Distribution of the most common bias types in CogniBias. Long-tail categories ensure diversity across 30+ psychological bias types.

F Dataset Access

The dataset and accompanying documentation are available at: `https://github.com/Mri1306/Cognitive-Bias-Dataset`

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: CogniBias is stated openly in the abstract and introduction as a benchmark dataset for cognitive biases in dialogue, which matches the contributions (Sections 1 and 3)

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: A separate "Discussion Limitations" section (Section 6) provides explicit notes on annotation ambiguity, coverage gaps, domain imbalance, as well as the reliance on LLM-generated dialogues.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include theoretical results, theorems, or proofs, since it is a dataset paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: Full details on dataset construction pipeline, annotation process, and analysis are provided in Section 3 (Dataset Design) and 4 (Dataset Analysis). Format information of the dataset is presented in Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset and annotation scripts will be released publicly with access instructions (Appendix E).

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: No training or model evaluation is reported, since the focus is dataset creation and analysis, not benchmarks.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not contain experiments that require statistical significance testing, only descriptive dataset statistics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

Justification: The work involves dataset construction and annotation, not computationally heavy experiments requiring resource reporting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: All dialogues were LLM-generated or expert-annotated. No personal data or privacy-sensitive content was collected. Ethical considerations are discussed in Section 7 (Broader Impact & Ethics).

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 7 (Broader Impact & Ethics) addresses both upsides (bias detection and fairness) and risks (misuse for persuasion).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The dataset is framed for evaluation only, not persuasion or manipulation. No personal or sensitive data is included (Section 7).

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Related benchmark datasets (e.g., WinoBias, StereoSet, CrowS-Pairs) are cited in Section 2 (Related Work). No external copyrighted data was reused.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: CogniBias is a new dataset, documented in Section 3 (Dataset Design), Section 4 (Analysis), and Appendix E (Dataset Access & Format)

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing was used. Annotations were performed by domain experts.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human-subject data collection was involved; all dialogues were generated with LLMs.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Section 3 (Dataset Design) explicitly describes that dialogues were generated using large language models and then annotated by experts.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.