

Deriving Necessary Validity Requirements for Evaluating the Consistency of LLM Responses

Alexander Ljapunov, Niklas Kiehne, Wolf-Tilo Balke

Institute for Information Systems
Technische Universität Braunschweig
Braunschweig, Germany

{ljapunov,kiehne,balke}@ifis.cs.tu-bs.de

Abstract

The probabilistic nature of large language models (LLMs) has regularly invited inquiries into their consistency with regard to repeated sampling, input variations, and conceptually related prompts. However, unlike accuracy benchmarks, consistency evaluations cannot rely on standardized measurement tools. We conduct a comprehensive survey of current methodologies evaluating language model consistency across 85 peer-reviewed publications, organizing them into a novel three-level taxonomy: (1) sampling, (2) semantic, and (3) relational consistency. Throughout this process, we identify three fundamental threats to the construct validity of LLM consistency measurements: Valid evaluations must disentangle intrinsic consistency from correctness, probabilistic uncertainty, and artifacts of constrained generation. Our systematic analysis shows that none of the 85 publications accounts for all three factors jointly, allowing construct-irrelevant variance to obscure our understanding of consistency.

1 Introduction

In everyday conversation, the responses we exchange are rarely arbitrary. Instead, we intuitively rely on an underlying consistency in the behavior of our interaction partner (Funder and Colvin, 1991). Continuous self-contradictions diminish trust in any person or system alike, regardless of any comparison against an established ground truth (Jang et al., 2022; Novikova et al., 2025). Therefore, it comes as no surprise that large language models (LLMs), as widely used text generation systems with an inherently probabilistic response process, naturally invite questions about their consistency (cf. Bender et al., 2021).

Continually, concerns have been raised regarding LLMs’ uncertainty under repeated sampling (Kuhn et al., 2023; Wang et al., 2023; Vashurin et al., 2025), sensitivity to prompt and order variations (Chang and Bergen, 2024; Wang et al., 2024b), and

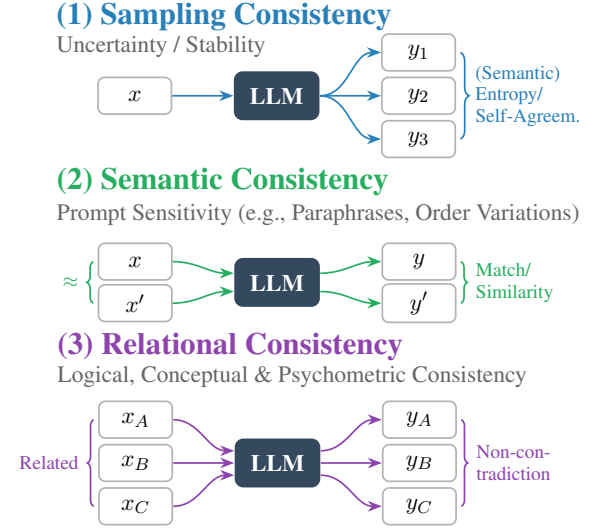


Figure 1: Overview of our proposed three-level taxonomy. It distinguishes between (1) sampling consistency (stability of multiple generations from one prompt), (2) semantic consistency (sensitivity to equivalent inputs), and (3) relational consistency (logical, conceptual, and psychometric relations between distinct inputs).

a general lack of a shared conceptual grounding across contexts (Nunes et al., 2024; Ding et al., 2025). Even without treating consistency as a normative requirement, understanding *in which way* and *to what extent* LLMs are inconsistent is vital for gaining insights into fundamental limitations, especially for domains with stringent predictability and safety demands. Similarly, unmeasured inconsistency naturally casts doubt on how broadly specific properties or findings from LLM research can actually be generalized (Jung et al., 2026).

Unlike classic performance benchmarks, consistency evaluations across responses cannot rely on standard, established metrics. Rather, authors generally need to design custom measurement frameworks in concordance with their research goals. While often sharing an intuition of underlying concepts – such as the need for comparisons between

outputs – there is no substantiated definition for what constitutes a valid measurement of language model (LM) consistency. We therefore ask:

*Are there **overarching necessary conditions** for a valid measurement of language model consistency?*

To answer this, we perform a comprehensive review of 85 peer-reviewed publications that conduct evaluations of the consistency of LM outputs. Our research process reveals how consistency measurements are regularly conflated with other model properties and artifacts, leading us to three basic, necessary requirements for construct validity, which, as our analysis shows, are not jointly satisfied by **any** evaluation within these 85 publications. Our contributions are:

- A meticulous **itemization and analysis** of current evaluation methodologies in **85 peer-reviewed publications** (Sections 3 and 5).
- A novel **three-level taxonomy** to categorize shared methodological concepts in the literature: (1) **sampling**, (2) **semantic**, and (3) **relational** consistency (Section 3).
- The identification of three primary sources of construct-irrelevant variance in current measurements, yielding corresponding **construct validity requirements** (Section 4). While not meant to be exhaustive, these form a foundational baseline to ensure evaluations measure consistency rather than experimental artifacts.

2 Language Model Consistency

LM consistency can broadly be understood as the semantic content¹ of responses not contradicting between generations. On the technical side, inconsistencies arise from both probabilistic uncertainty, where input variations or sampling noise trigger path-dependent divergences (Xu et al., 2024), and under-generalization, where models fail to transfer conceptual meaning across contexts (Chang and Bergen, 2024; Nunes et al., 2024). Both can steer models into conflicting areas of their heterogeneous training data (Xu et al., 2024) or into arbitrary hallucinations (Manakul et al., 2023; Kalai et al., 2025).

We strictly focus on *internal* consistency *across* distinct outputs.² This distinguishes our scope from related constructs: (1) *robustness*, or ground-truth

¹Often termed "belief" in the literature (e.g., Li et al., 2019; Jang et al., 2022; Hase et al., 2023; Scherrer et al., 2023).

²We avoid the synonymous term "self-consistency" (e.g., Mitchell et al., 2022; Chen et al., 2023) to prevent confusion with the prominent prompting technique (Wang et al., 2023).

accuracy against adversarial shifts (Wang et al., 2022; Mizrahi et al., 2024); (2) *external/factual consistency*, representing faithfulness to external knowledge (Honovich et al., 2022; Jang et al., 2022); and (3) *intra-output coherence* within a single context window, such as reasoning faithfulness (Lanham et al., 2023; Paul et al., 2024).

Understanding consistency is crucial for interpreting LM research: From a psychometric perspective (American Educational Research Association et al., 2014), constructs cannot be validly attributed if behavior shifts arbitrarily across settings or samples. Transferring standard reliability measures to LLM evaluation (cf. Löhn et al., 2024) – namely test-retest (repeated sampling), alternate forms (rephrased inputs), and internal consistency (relational checks) – forms the conceptual foundation for our proposed three-level taxonomy.

3 A Three-Level Taxonomy of Language Model Consistency

To gain an extensive overview of methodologies used for evaluating language model consistency, we conducted a targeted literature search process. Papers were filtered in three steps: (1) constructing a candidate pool via *targeted keyword searches* on the ACL Anthology as well as DBLP and Google Scholar, which, after completion of this initial acquisition, we cross-referenced with a recent extensive survey on language model consistency (Novikova et al., 2025) to find potential blind spots; (2) selecting only *peer-reviewed papers* from this pool, published either in ACL conferences and journals or in other highly reputable venues (such as NeurIPS, ICLR, ICML, COLM, AIES, TMLR, or Nature); and (3) filtering on whether they conduct an *actual evaluation* of LM consistency, explicitly excluding purely theoretical or methodological works lacking concrete evaluation from our structured literature review. Our final literature pool consisted of 85 publications, each of which we systematically analyzed regarding the respective evaluation methodology.

To categorize these methods, we propose a hierarchical, three-level taxonomy based on inputs and constraints. Prior taxonomies have relied on formal definitions (e.g., semantic, logical, and factual, cf. Jang et al., 2022, or negation, implication, transitivity, and factual, as well as their composition, cf. Cheng et al., 2025) that fail to capture large parts of recent research. Novikova et al. (2025) inherit

the above definitions from Jang et al. (2022) and expand them with a broader umbrella term of informal consistency. Notably, none of the above considers inconsistency from sampling. Here, the taxonomy by Liang et al. (2024) has an orthogonal focus on sampling diversity for inference-time methods, while mostly ignoring input variations.

Inspired by Scherrer et al. (2023), our three-level framework separates variations caused by sampling from those caused by input differences, and, like Jang et al. (2022), distinguishes consistency across semantically equivalent inputs from consistency regarding more complex relations:

1. **Sampling Consistency:** Non-contradiction of multiple sampled outputs for the exact same input (like Liang et al., 2024).
2. **Semantic Consistency:** Consistency of outputs for semantically equivalent inputs, both through formal semantic checks (cf. Jang et al., 2022) and informal heuristics (cf. Novikova et al., 2025).
3. **Relational Consistency:** Non-contradiction across distinct, related inputs, capturing formal logical constraints (as in Jang et al., 2022; Cheng et al., 2025) as well as implicit conceptual relations and psychometric correlations.

As discussed in Section 2, we explicitly exclude factual consistency. Furthermore, while these levels logically nest within one another, we conceptualize them as separate from a methodological standpoint.

A detailed itemization of the 85 publications – which may appear across multiple levels – is provided in Table 1 (sampling) as well as Tables 3 and 4 in Appendix B (semantic and relational). Each entry outlines the evaluated **CONSISTENCY CHECKS** and **METRIC**; **OUTPUT** indicates the response format: Direct answers (i.e., any commitment in the first token) in normal text, ***bold italics*** for chain-of-thought (CoT), and **bold** for open-ended outputs requiring heuristic extraction, annotated by extraction method, e.g., man (manual), NLI (natural language inference models), emb (embedding similarity), CE (cross-encoder similarity), LLM (LLM extraction). Multiple heuristics are marked with *, with further details about the methods provided in the **METRIC** column.

3.1 Sampling Consistency

Beyond measuring output stability, sampling consistency underpins uncertainty quantification (e.g.,

Vashurin et al., 2025), capable of predicting robustness (Rabinovich et al., 2023; Jiang et al., 2025) and hallucinations (Farquhar et al., 2024; Nikitin et al., 2024). Furthermore, sampling variance is a widely used signal to improve reasoning, both during inference (e.g., Self-Consistency prompting; Wang et al., 2023; see Liang et al., 2024 for an overview) and training, increasing the sampling consistency of correct reasoning traces (Yue et al., 2025) in reinforcement learning algorithms like GRPO (Shao et al., 2024).

Table 1 provides an overview of 18 papers evaluating sampling consistency, half of which focus exclusively on this level, with the rest measuring sampling and higher-level consistency in parallel.

Consistency Checks Most works focus on a simple uncertainty or robustness evaluation on factual question answering (QA) or natural language understanding (NLU) benchmarks. However, three address uncertainty in subjective domains like political or opinion questions (Ceron et al., 2024; Durmus et al., 2024), with Scherrer et al. (2023) crucially finding that LMs exhibit significantly higher uncertainty in high-ambiguity moral scenarios compared to low-ambiguity ones with correct choices.

Metrics & Outputs Metrics are highly contingent on output formats. To simplify the calculation of self-agreement or entropy, 8 of 18 papers sample direct, single-word answers or measure first-token probabilities. However, Li et al. (2024a) note that such constraints can obstruct natural model behavior, observing that unconstrained generation alters measured answer variance. Here, uncertainty quantification evaluates open-ended generation by semantically clustering responses (e.g., via NLI) to avoid penalizing rewording (e.g., Kuhn et al., 2023; Vashurin et al., 2025; see Table 1 for more approaches). Interestingly, while effective here, similar semantic approaches are rarely used in higher-level consistency evaluations.

3.2 Semantic Consistency

Evaluating semantically equivalent inputs tests whether models generalize an understanding of meaning rather than superficially pattern-match form (Ohmer et al., 2024; Chang and Bergen, 2024). Such sensitivity evaluations are powerful as local soundness (Fluri et al., 2024) or alternate-forms reliability checks (Shu et al., 2024; Ceron et al., 2024), but are no indicator of global coherence.

PAPER (<i>Venue</i>)	CONSISTENCY CHECKS	METRIC	OUTPUT
Kuhn et al. (2023) (<i>ICLR</i>)	Semant. equiv. sampled open-ended outputs (determ. by NLI)	Entropy of output equiv. classes	Open-ended ^{NLI}
Scherrer et al. (2023) (<i>NeurIPS</i>)	Distr. of 10 sampl. actions in both unambig. & ambig. moral scenarios	Entropy of answ. distr.	Binary-choice
Ohmer et al. (2023) (<i>GEM</i>)	Same answ. if sampled twice	% of cons. pairs	Yes/No
Ceron et al. (2024) (<i>TACL</i>)	Stat. signific. of 30 sampled polit. opinion answ. (test-retest reliab.)	95% conf. interval of 1000-rep. bootstrap	Agree/Disagree
Lin et al. (2024) (<i>TMLR</i>)	Semant. equiv. sampled open-ended outputs (determ. by NLI)	Diff. metrics from NLI-induc. adjacency matrix	Open-ended ^{NLI}
Li et al. (2024a) (<i>LREC-COLING</i>)	Ratio of same answ. for 5 samples	# of majority answer + # of other options - 1	MCQ; open-ended ^{man}
Farquhar et al. (2024) (<i>Nature</i>)	Semant. equiv. sampled open-ended outputs (determ. by NLI)	Entropy of output equiv. classes	Open-ended ^{NLI}
Zhang et al. (2024a) (<i>NAACL</i>)	Distrib. of all sampled answ.	Entropy of answer distr.	Single-word
Wang et al. (2024a) (<i>ACL</i>)	Self-agreem. of binary LLM as judge preferences	Entropy of pref. responses	Pref. scores; CoT + Scores
Durmus et al. (2024) (<i>COLM</i>)	Token distr. on opinion quest.	Token distribution	MCQ tkn. prob.
Nikitin et al. (2024) (<i>NeurIPS</i>)	Semant. equiv. sampled open-ended outputs (determ. by NLI)	von Neumann Entropy of sem. graph kernels	Open-ended ^{NLI}
Ye et al. (2024) (<i>NeurIPS Datas. & Benchm. Track</i>)	Token/answ. distrib. of different classes	Set Size: # of class labels above probab. thresh.	MCQ tkn. prob.
Ohmer et al. (2024) (<i>CL</i>)	Same answ. if sampled twice	% of consis. pairs	Yes/No, 1-word
Lee et al. (2025) (<i>COLING</i>)	Cons. sampled scores in text evaluation	Krippendorff's α	Num. & Likert Scales, Yes/No
Liu et al. (2025) (<i>ICML</i>)	Self-agr. of LLM-as-a-judge pref.	% of majority answer	(CoT) bina.-ch.
Vashurin et al. (2025) (<i>NeurIPS</i>)	Open-e. answer sem. equiv. to other samples (determ. by cross-encoder)	1—prob. of answer seq. & avg. 1—sim. to samples	Open-ended ^{CE}
Fadeeva et al. (2026) (<i>ICLR</i>)	Open-e. answer sem. equiv. to likely beam search outp. (det. by NLI)	Like Vashurin et al. (2025) & Lin et al. (2024) for beams	Short-form answers ^{NLI}
Cao et al. (2026) (<i>EACL</i>)	Sem. equiv. tokens in each decoding step (det. by clustering of emb.)	Product of cluster prob. masses for each token in answer seq.	Open-ended ^{emb}

Table 1: Overview of methods in sampling consistency evaluations. **Bold** text in the **CONSISTENCY CHECKS** column indicates tasks without a ground truth (**R1**). **Bold italic** text in the **OUTPUT** column represents outputs with CoT, **bold** text represents open-ended generation with the extraction heuristic provided in superscript (**R3**).

We provide a detailed itemization of 51 papers evaluating semantic consistency in Table 3 in Appendix B. These works test sensitivity of models to prompt variations such as paraphrases, translations, negations of binary questions, option reordering, or instruction and output format variations.

Consistency Checks Most works test robustness deficiencies on NLU or QA tasks, with early influential work (Jiang et al., 2020; Elazar et al., 2021) spawning numerous evaluations of prompt sensitivity to paraphrased or translated queries for knowledge base (KB) or graph (KG) facts. As with sampling, some evaluate input variations in subjective or ambiguous domains like morality, personality, or politics (e.g., Agarwal et al., 2024; Röttger et al., 2024). Crucially, comparisons show that models can achieve very high consistency on tasks with an unambiguously correct choice, but exhibit pronounced belief shifts (e.g., in political opinions) across translated or paraphrased inputs of higher-

ambiguity tasks (Scherrer et al., 2023; Zhou and Zhang, 2024; Moore et al., 2024).

Metrics The vast majority of the field (45 out of 51 papers) follows works such as Elazar et al. (2021), relying on deterministic point-wise metrics, computing the raw percentage of exact matches or the average of some similarity measure between outputs. A small minority accounts for the fact that discrete mismatches can simply stem from sampling noise. To capture actual inter-prompt belief shifts, these studies quantify consistency probabilistically through either the (e.g., Jensen-Shannon) divergence (JSD) between Monte-Carlo estimated answer (e.g., Scherrer et al., 2023) or token logit (e.g., Moore et al., 2024) distributions, or by comparison of token rankings (e.g., Qi et al., 2023).

Outputs Most evaluations (30 out of 51) artificially restrict to forcing multiple-choice (MCQ) or otherwise direct answers. 6 additionally allow for CoT before answer commitment. 15 publi-

cations evaluate open-ended outputs, relying on NLI, semantic similarity, or LLM-based heuristics for consistency comparisons (e.g., Honovich et al., 2022; Zhao et al., 2024) or answer extraction (e.g., Moore et al., 2024). Notably, papers directly comparing format constraints find a critical inconsistency: LLM behavior shifts drastically when forced to predict a single next token. Open-ended answers often yield vastly different distributions (Röttger et al., 2024) and are generally far more robust and less biased toward option-order artifacts than their forced-choice counterparts (Wang et al., 2024b,c).

3.3 Relational Consistency

Relational consistency demands non-contradiction between answers to related inputs. Given the vastness of possible relationships, evaluations typically target specific local connections. As detailed for the 40 papers in Table 4 in Appendix B, these evaluations broadly fall into three categories: hand-crafted **logical constraints**, distinct but **conceptually related** contexts, and **psychometric** relations.

Logical consistency Logical consistency examines outputs’ adherence to strict, a priori defined constraints, typically using propositional questions or probability forecasts (e.g., Paleka et al., 2025). Most commonly studied is transitivity for NLI tasks, QA, and preference orderability. Other prominent constraints include implication (e.g., of ontological relations in KGs) and monotonicity (e.g., Kassner et al., 2021; Ross et al., 2024), as well as con- and disjunctions of input prompts.

Conceptual consistency While modern LLMs handle local semantic and logical constraints decently, they often fail to generalize across distinct but conceptually highly related contexts, exhibiting a lack of "concept mastery" (Nunes et al., 2024). This inconsistency can generally be seen in behavioral **gaps**: generator-validator (GV) inconsistency (failing to validate their own generations) (Li et al., 2024b), the safety discrimination-generation gap (Ding et al., 2025), the reversal curse (Berglund et al., 2024), and hard-to-easy inconsistency (Yang et al., 2024b). Similarly, abstract value judgments do not align with concrete actions (Jain et al., 2024; Shen et al., 2025), an effect more pronounced if there is no a priori correct answer (Xu et al., 2025).

Psychometric consistency Here, softer notions of agreement or correlations are evaluated. While LLMs show high inter-question reliability within a

single human-validated test (Serapio-García et al., 2025; Ye et al., 2025), akin to conceptual consistency, they lack global cross-test correlations. Unlike humans, their abstract moral values do not correlate with their responses to concrete scenarios (Nunes et al., 2024), and neither do questionnaire answers and behavior in downstream tasks (Jung et al., 2026). Political or opinion stances also do not maintain alignment with the same demographic groups across related topics (Santurkar et al., 2023; Durmus et al., 2024; Ceron et al., 2024).

Metrics 33 out of 40 publications rely on single-sample metrics, such as ratios of logical constraint violations (commonly building on Li et al., 2019), simple ratios of corresponding answer pairs for a conceptual evaluation, and psychometric measures on answers (e.g., Cronbach’s α). Only a minority of works conduct probabilistic comparisons. Given the prevalence of MCQ settings, these typically compare first-token probabilities (Santurkar et al., 2023; Moore et al., 2024), with results by Xu et al. (2025) suggesting that single-sample metrics significantly overestimate answer shifts compared to underlying token distributions. As a rare example of a more complex probabilistic constraint, Yang et al. (2024b) use Monte-Carlo sampling to estimate the probability of solving an easy question conditioned on solving a strictly harder one.

Outputs 26 of 40 papers exclusively force direct answers (e.g., Boolean, numeric, MCQ), with 10 recent works utilizing CoT prior to extraction. Only 4 works evaluate open-ended outputs, relying on complex methods like value extraction with multi-agent pipelines (Ye et al., 2025) or iterative checking of error messages when translating between natural language and code (Min et al., 2024).

4 Construct Validity Requirements for Consistency Evaluations

When quantifying consistency, evaluations must distinguish the target construct from other sources of observed agreement. High agreement can simply follow from high correctness, inter-input disagreement may reflect sampling uncertainty rather than be caused by input differences, and constraints in response generation can induce either pattern. Without accounting for these factors, observed consistency remains ambiguous in interpretation.

In line with the closely related, broader recommendations for LLM evaluation of Bean et al.

(2026), we identify three necessary requirements for **construct validity**: disentangling consistency from correctness (**R1**), sampling uncertainty (**R2**), and constrained generation artifacts (**R3**). We note that these requirements form a necessary but non-exhaustive baseline for claims about consistency as a standalone construct. They do not preclude narrower claims conditional on a particular task, decoding method, or response format.

4.1 R1: Disentangle Consistency from Correctness

Observed consistency on correctness-based tasks can be misleading. Since language models are optimized for accuracy and truthfulness, consistently correct outputs may stem from independent inferences rather than a coherent understanding of underlying concepts (Ohmer et al., 2024). As a valid ground truth is contradiction-free, the possible disagreement is upper bound by the sum of error rates. Even assuming independent errors, an LLM’s predictions for two prompt paraphrases that are each 90% accurate on a binary task have an expected agreement of $0.9^2 + 0.1^2 = 82\%$. High agreement can arise from accuracy alone without intrinsic cross-prompt consistency. Rather, in correctness-bound tasks, consistency is often merely an incomplete predictor of accuracy itself (e.g., Rabinovich et al., 2023; Ye et al., 2024). Paradoxically, an LLM producing systematically *incorrect* outputs would be a stronger indicator of model consistency (cf. Tan et al., 2025; Ganesh et al., 2026).

To establish that consistency captures more than correctness, evaluations should include inputs without definitive ground truths. Consistency is a natural measure in domains where responses can be individually plausible but collectively contradictory. Unambiguous tasks might still serve as a control to establish a baseline of inconsistency caused by technical artifacts (i.e., a lack of robustness). Controlling for such deficiencies can isolate tendencies of actual conceptual belief shifts for, e.g., ambiguous moral dilemmas or controversial topics (cf. Scherrer et al., 2023; Zhou and Zhang, 2024).

4.2 R2: Disentangle Consistency from Uncertainty

Standard evaluations relying on single-sample outputs conflate cross-prompt inconsistency with inherent probabilistic uncertainty. While deterministic decoding is often assumed to eliminate this randomness, it merely masks the model’s internal

margins. For instance, contradicting greedy decodings across two equivalent prompts might reflect minimal shifts around a 50/50 decision boundary, whereas point estimates can conversely obscure a drop in confidence (e.g., from 95% to 55%) if the top output remains unchanged. Without examining probabilities, random variation is misdiagnosed as a failure to generalize, an artifact that will converge to biased error rates rather than averaging out.

To isolate actual probabilistic belief shifts, sampling uncertainty must be treated as the null hypothesis (Ohmer et al., 2024). Consistency evaluations should therefore measure divergence between output distributions, estimated via Monte-Carlo sampling or token logits. Modeling LM beliefs probabilistically (e.g., Scherrer et al., 2023) and applying specialized probabilistic constraints allow for explicitly factoring out sampling noise and validly measuring inter-input inconsistencies.

4.3 R3: Disentangle Consistency from Constrained Generation Artifacts

The observed consistency of an LLM is fundamentally altered by the generation format used in the evaluation. While forcing exactly matchable direct outputs (e.g., "Yes/No", MCQ, or Likert scales) facilitates sound evaluation and probabilistic modeling, it artificially restricts the model’s underlying autoregressive reasoning process. Given standard formal assumptions, single-forward-pass predictions are bottlenecked by theoretical limits of transformer expressivity (Merrill and Sabharwal, 2023); premature first-token commitments frequently cause hallucinations that snowball for the subsequent generation process (Zhang et al., 2024b). They significantly increase selection and position biases (Zhou et al., 2024) as well as prompt sensitivity (Röttger et al., 2024; Wang et al., 2024b), while overestimating certainty compared to the accumulated noise in sequential generation (Wang et al., 2024c; Xie et al., 2024).

For autoregressive language models, avoiding conflation with these restricted output artifacts requires allowing extended generation, bypassing such computational bottlenecks (Merrill and Sabharwal, 2024; Jiang et al., 2026). CoT prompting (and native reasoning models) has the advantage of mitigating reasoning limits while typically concluding with an output that allows rule-based extraction. In contrast, open-ended generation provides the most ecologically valid measure of natural model behavior. However, evaluating open-ended text ne-

R1: Correctness			R2: Uncertainty		R3: Extended Generation	
R1	R2	R3	#	Papers		
Sampling Consistency						
✗	—	✗	5	Ohmer et al. (2023) • Zhang et al. (2024a) • Ye et al. (2024) • Ohmer et al. (2024) • Lee et al. (2025)		
✗	—	✓	10	Kuhn et al. (2023) • Lin et al. (2024) • Li et al. (2024a) • Farquhar et al. (2024) • Wang et al. (2024a) • Nikitin et al. (2024) • Liu et al. (2025) • Vashurin et al. (2025) • Fadeeva et al. (2026) • Cao et al. (2026)		
✓	—	✗	3	Scherrer et al. (2023) • Ceron et al. (2024) • Durmus et al. (2024)		
✓	—	✓	0	—		
Semantic Consistency						
✗	✗	✗	20	Li et al. (2019) • Asai and Hajishirzi (2020) • Jiang et al. (2020) • Elazar et al. (2021) • Newman et al. (2022) • Fierro and Søgaard (2022) • Jang et al. (2022) • Albrecht et al. (2022) • Zhou et al. (2022) • Hase et al. (2023) • Hagström et al. (2023) • Jang and Lukasiewicz (2023) • Ohmer et al. (2023) • Rabinovich et al. (2023) • Zhao et al. (2024) • Mizrahi et al. (2024) • Rajan et al. (2024) • Ohmer et al. (2024) • Lee et al. (2025) • Ganesh et al. (2026)		
✗	✗	✓	12	Honovich et al. (2022) • Raj et al. (2022) • Li et al. (2024a) • Yang et al. (2024a) • Wang et al. (2024a) • Wang et al. (2024b) • Raj et al. (2024) • Calanzone et al. (2025) • Ghosh et al. (2025) • Liu et al. (2025) • Jang et al. (2025) • Bu et al. (2026)		
✗	✓	✗	2	Kassner and Schütze (2020) • Qi et al. (2023)		
✗	✓	✓	0	—		
✓	✗	✗	5	Kiehne et al. (2022) • Kiehne et al. (2024) • Shu et al. (2024) • Ceron et al. (2024) • Jung et al. (2026)		
✓	✗	✓	8	Fluri et al. (2024) • Bonagiri et al. (2024) • Röttger et al. (2024) • Wang et al. (2024c) • Zhou and Zhang (2024) • Zhou et al. (2024) • Paleka et al. (2025) • Liu et al. (2026)		
✓	✓	✗	3	Scherrer et al. (2023) • Agarwal et al. (2024) • Durmus et al. (2024)		
✓	✓	✓	1*	(*) Moore et al. (2024)		
Relational Consistency						
✗	✗	✗	14	Li et al. (2019) • Ribeiro et al. (2019) • Asai and Hajishirzi (2020) • Kassner et al. (2021) • Jang et al. (2022) • Mitchell et al. (2022) • Hase et al. (2023) • Gu et al. (2023) • Chen et al. (2023) • Berglund et al. (2024) • Li et al. (2024b) • Jang and Lukasiewicz (2023) • Kassner et al. (2023) • Rajan et al. (2024)		
✗	✗	✓	8	Min et al. (2024) • Zhang et al. (2024b) • Jiang et al. (2024) • Calanzone et al. (2025) • Ghosh et al. (2025) • Liu et al. (2025) • Bajpai and Chakraborty (2025) • Ding et al. (2025)		
✗	✓	✗	1	Yang et al. (2024b)		
✗	✓	✓	0	—		
✓	✗	✗	6	Nye et al. (2021) • Jin et al. (2022) • Ceron et al. (2024) • Jain et al. (2024) • Shen et al. (2025) • Serapio-García et al. (2025)		
✓	✗	✓	5	Fluri et al. (2024) • Ross et al. (2024) • Paleka et al. (2025) • Ye et al. (2025) • Jung et al. (2026)		
✓	✓	✗	5	Santurkar et al. (2023) • Durmus et al. (2024) • Nunes et al. (2024) • Moore et al. (2024) • Lehr et al. (2025)		
✓	✓	✓	1*	(*) Xu et al. (2025)		

Table 2: Analysis of consistency evaluation methodologies across 85 unique publications based on the validity requirements (R1-R3) defined in Section 4. ✓ indicates the requirement is satisfied, ✗ indicates it is not accounted for, and – denotes the requirement is not applicable. Notably, evaluations marked with an asterisk (*) satisfy all three criteria across separate experiments, but fail to satisfy them jointly within a single measurement.

cessitates post-hoc extraction or comparisons (e.g., via NLI or LLMs), which might introduce their own second-order validity and reliability concerns. For example, Wu et al. (2025) find that such heuristic consistency measurements can align poorly with semantic-similarity judgments of human raters.

5 Analyzing the Methodological Landscape in LM Consistency

We systematically code the methodologies of all 85 publications against our three validity criteria (R1-R3) in Table 2. Evaluations receive a ✓ if they: (R1) test ambiguous or subjective domains without a priori correct answers;³ (R2) disentangle uncertainty using probabilistic comparisons rather than single-sample matches;⁴ and (R3) allow for any generation beyond direct commitments (either

through CoT or open-ended generation). ✗ indicates that the evaluation does not account for the factor. If a paper features multiple evaluations, we prioritize the assignment of ✓ over ✗. We mark R2 with – in sampling consistency as the requirement does not apply. We provide more details on the coding in Appendix A.

The most striking finding from our review is that not a single publication satisfies all three validity requirements simultaneously. Instead, evaluations frequently sacrifice one dimension of construct validity in satisfying another. Even Moore et al. (2024) and Xu et al. (2025) only consider all three requirements in separate experiments, but not *jointly* within a single measurement.

Holdovers from Masked Language Models Evaluations frequently fail R2 or R3 due to a reliance on outdated evaluation paradigms. Early influential works (e.g., Li et al., 2019; Jiang et al., 2020; Elazar et al., 2021) established consistency

³If ✓ is assigned, we also highlight the data or task in **bold** in the **CONSISTENCY CHECKS** column in Tables 1, 3, and 4.

⁴Marked **bold** in the **METRIC** column in Tables 3 and 4.

measures tailored to masked language models like BERT (Devlin et al., 2019), where single-forward-pass matches were the ecologically valid standard. As the field transitioned to autoregressive LLMs, researchers often iterated upon these frameworks rather than formalizing setups for probabilistic, sequential generation from scratch.

Consequently, modern semantic and relational evaluations exhibit a stark mutual exclusivity between uncertainty disentanglement (**R2**) and unconstrained formats (**R3**). Studies often consider the role of the output format for consistency (e.g., Röttger et al., 2024; Wang et al., 2024c) but fall back on simplistic, single-sample matches (satisfying **R3** but failing **R2**), or formulate rigorous probabilistic frameworks (e.g., Durmus et al., 2024; Yang et al., 2024b) that treat LLMs like simple classifiers by sampling direct outputs or comparing the first-token distribution (satisfying **R2** but failing **R3**). In fact, **R2** is especially rare, met by only 7 of 51 semantic and 7 of 40 relational consistency publications. Even comprehensive evaluations (Moore et al., 2024; Xu et al., 2025) exhibit this limitation, computing probabilistic divergence exclusively over first-token logits and reverting to point-wise matches for CoT and open-ended text.⁵

This is a notable methodological disconnect across taxonomy levels: While the sampling consistency literature has developed sophisticated uncertainty quantification frameworks specifically for unconstrained text, similar approaches remain mostly absent from semantic and relational evaluations.

Methodological Heterogeneity Furthermore, outside of specific niches like moral alignment or psychometric studies, the majority of semantic, logical, and conceptual evaluations violate **R1** by relying strictly on ground-truth accuracy tasks. Even at the sampling level, where **R2** is not required, only 3 of 18 publications evaluate subjective domains, and, notably, none of these allows for CoT or open-ended generation (**R3**) despite the breadth of approaches in uncertainty quantification at the sampling consistency level. Echoing sentiments voiced by Novikova et al. (2025), we find current evaluation methodologies fractured and lacking valid measurement frameworks. We stress that our coding of the publications against the criteria was intentionally lenient. While future work may identify additional conditions

for validity, failing to isolate a measurement from these basic artifacts already compromises its interpretation with regard to LM consistency. **No study** in our corpus manages to satisfy even this foundational baseline.

6 Discussion

Do we understand how consistent LLMs are?

While findings in the literature clearly demonstrate that LLMs exhibit inconsistencies, the conflation of different model properties makes it difficult to precisely isolate how and to what extent underlying conceptual inconsistencies occur. Evaluations that violate **R1** may hold undeniable practical value as rigorous and valid proxies for assessing model robustness, uncertainty, or capability transfer, yet fail to deepen our understanding of consistency as a standalone property. Furthermore, because no study jointly addresses sampling uncertainty (**R2**) alongside unconstrained generation (**R3**), we know surprisingly little about LLM behavior with regard to cross-prompt consistency in natural, autoregressive settings. We want to stress that our analysis is not meant to be read as a critique of individual works. Rather, failing to control for all three factors limits the generalizability of a result to any overarching claim about quantifying LLM response consistency itself. Until evaluations jointly address these factors, LLM consistency as a property remains beyond our grasp.

The Cost of Valid Evaluation We theorize that the striking finding of no study jointly covering **R2** and **R3** in part reflects the computational and practical costs of addressing both requirements simultaneously. While uncertainty disentanglement can be achieved in a single forward pass per input by simply comparing output token logits, forcing single-token outputs violates **R3**. Conversely, extended generation already requires a non-trivial increase in computational resources. Uncertainty estimation for unconstrained text is often restricted to methods such as Monte-Carlo estimation, multiplying inference costs by the number of output samples. Furthermore, probabilistic comparisons, especially beyond first-token probabilities, entail a significant theoretical formalization overhead, particularly for already complex logical or conceptual consistency constraints. We acknowledge that for many researchers, point-wise comparisons and forced direct answers simply remain the most accessible options.

⁵Moore et al. (2024) only consider open-ended text in their semantic, not in their relational consistency evaluation.

Jointly Operationalizing R1-R3 The three requirements constrain different evaluation stages: **R1** concerns the selection of the evaluation domain and interpretation of agreement metrics, **R3** concerns response elicitation, and **R2** concerns aggregation and comparison. None logically implies another. Their concrete joint implementation, especially regarding **R2**, will depend on the evaluated consistency relation, which varies across and within taxonomy levels. Nevertheless, here, we want to provide a possible schematic to exemplify how the requirements could jointly be operationalized within a single evaluation.

For semantic consistency, an evaluation could apply prompt variations to ambiguous inputs (**R1**), sample multiple CoT or open-ended responses per variation (**R3**), and map them to either predefined choices or semantic equivalence classes (as in Kuhn et al., 2023; Nikitin et al., 2024), yielding empirical response distributions. These can be used to measure a between-paraphrase divergence (e.g., JSD) relative to within-prompt sampling uncertainty (**R2**). Notably, Scherrer et al. (2023) already apply a similar distributional framework for predefined direct-answer choices in moral scenarios, and adding CoT or open-ended responses would then additionally address **R3**.

While **R1** and **R3** transfer similarly to relational consistency, **R2** requires relation-specific probabilistic formalization. Logical evaluations might estimate assent probabilities through sampling and measure violations against probabilistic constraints, such as the ones applied by Paleka et al. (2025). Translating **R2** to conceptual gaps like generator-validator consistency would mean comparing the generator’s distribution of answers with the validator’s preference distribution. In psychometric evaluations, repeated generations could, for example, conceptually be treated similarly to a population of respondents (cf. Löhn et al., 2024).

Arbitrariness or Pluralism? Beyond these methodological questions, a deeper, more abstract interpretive ambiguity concerns the unit of analysis. Consistency evaluations that aggregate responses at the model level implicitly treat the LLM as a single entity expected to express a unified viewpoint. Yet, from the perspective of LLM psychometrics, as Löhn et al. (2024) argue, models are not well-suited to such an individual-level interpretation: Their response distributions often resemble heterogeneous response populations. Particularly in subjective do-

main, interpreting disagreement is ambiguous. It may arise from arbitrary variance, sensitivity, or hallucinations, or it may reflect a (potentially desirable) plurality of internally coherent viewpoints (cf. Sorensen et al., 2024). Stenseke et al. (2026), for example, have directly raised this concern specifically for moral consistency, arguing that justifiable inconsistencies between reasonable responses should not necessarily be treated as a defect. Nevertheless, quantification of model-level consistency remains relevant for aspects such as predictability and safety. An avenue for future evaluations could therefore be to actively distinguish between intra-viewpoint inconsistency and inter-viewpoint heterogeneity. Analogous to how **R2** separates intra-prompt uncertainty from cross-prompt shifts, this could help conceptually disentangle arbitrary contradictions from pluralistic diversity.

7 Conclusion

Through a comprehensive review of 85 peer-reviewed publications, we introduced a three-level taxonomy – (1) sampling, (2) semantic, and (3) relational consistency – to organize the landscape of language model consistency evaluations. We identified three prevalent threats to construct validity: Evaluations fail to disentangle consistency from correctness (**R1**), probabilistic uncertainty (**R2**), or constrained generation artifacts (**R3**).

Strikingly, zero evaluations in our analysis can satisfy all three validity requirements simultaneously. Studies frequently conflate a model’s ability to give the correct answer with consistency, and evaluation frameworks are still regularly built upon outdated, point-wise metrics and forced direct outputs designed for older language models, rather than being constructed with the uncertain, autoregressive nature of modern LLMs in mind. While the NLP community widely acknowledges that LLMs exhibit inconsistencies, the understanding of their actual extent and nature remains obscured by these methodological artifacts.

Limitations

We recognize that our search process might have overlooked either conceptually highly related publications simply not captured by our searches or a few select very recent publications. On this note, our literature survey process was strictly limited to peer-reviewed publications from reputable venues to ensure our assessment covered published

works meeting a standard of methodological quality. However, given the rapid pace of LLM and AI research, this will miss very recent evaluation paradigms that currently exist only as preprints.

Additionally, our focus lies only on the evaluation of consistency, meaning literature exclusively covering methods for improving consistency, inference-time or training based on consistency research, or theoretical work are not categorized as part of our survey.

Furthermore, our coding in Section 5 was deliberately simplified from a more nuanced assessment to a binarized one, as almost all papers fall into one of two camps per requirement. Consequently, we acknowledge that our analysis misses a small handful of works such as [Ohmer et al. \(2024\)](#), who provide notable theoretical rigor in deliberating about the role of correctness as an upper bound (**R1**) and sampling as a baseline (**R2**) for semantic consistency. However, their concrete practical methodology falls short of our requirements.

References

- Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. [Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.
- Josh Albrecht, Ellie Kitanidis, and Abraham Fetterman. 2022. Despite "super-human" performance, current LLMs are unsuited for decisions about ethics and safety. In *NeurIPS ML Safety Workshop*.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.
- Akari Asai and Hannaneh Hajishirzi. 2020. [Logic-guided data augmentation and regularization for consistent question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650, Online. Association for Computational Linguistics.
- Ashutosh Bajpai and Tanmoy Chakraborty. 2025. [Temporal Referential Consistency: Do LLMs Favor Sequences Over Absolute Time References?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17618–17636, Suzhou, China. Association for Computational Linguistics.
- Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, and 23 others. 2026. [Measuring what matters: Construct validity in large language model benchmarks](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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 '21*, pages 610–623, New York, NY, USA. Association for Computing Machinery.
- Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. [The reversal curse: LLMs trained on "a is b" fail to learn "b is a"](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Vamshi Krishna Bonagiri, Sreeram Vennam, Priyanshu Govil, Ponnurangam Kumaraguru, and Manas Gaur. 2024. [SaGE: Evaluating moral consistency in large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14272–14284, Torino, Italia. ELRA and ICCL.
- Yuyan Bu, Xiaohao Liu, ZhaoXing Ren, Yaodong Yang, and Juntao Dai. 2026. Align Once, Benefit Multilingually: Enforcing Multilingual Consistency for LLM Safety Alignment. In *The Fourteenth International Conference on Learning Representations*.
- Diego Calanzone, Stefano Teso, and Antonio Vergari. 2025. Logically Consistent Language Models via Neuro-Symbolic Integration. In *The Thirteenth International Conference on Learning Representations*.
- Qi Cao, Andrew Gambardella, Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. 2026. [Semantic Token Clustering for Efficient Uncertainty Quantification in Large Language Models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–696, Rabat, Morocco. Association for Computational Linguistics.
- Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. [Beyond Prompt Brittleness: Evaluating the Reliability and Consistency of Political Worldviews in LLMs](#). *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

- Tyler A. Chang and Benjamin K. Bergen. 2024. [Language model behavior: A comprehensive survey](#). *Computational Linguistics*, 50(1):293–350.
- Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R. Bowman, and Kyunghyun Cho. 2023. Two Failures of Self-Consistency in the Multi-Step Reasoning of LLMs. *Transactions on Machine Learning Research*.
- Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. 2025. [Empowering LLMs with Logical Reasoning: A Comprehensive Survey](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Peng Ding, Wen Sun, Dailin Li, Wei Zou, Jiaming Wang, Jiajun Chen, and Shujian Huang. 2025. [SDGO: Self-Discrimination-Guided Optimization for Consistent Safety in Large Language Models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5023–5037, Suzhou, China. Association for Computational Linguistics.
- Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askeil, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. In *First Conference on Language Modeling*.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhishava Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. [Measuring and improving consistency in pretrained language models](#). *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Ekaterina Fadeeva, Maiya Goloburda, Aleksandr Rubashevskii, Roman Vashurin, Artem Shelmanov, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. 2026. Don’t Throw Away Your Beams: Improving Consistency-based Uncertainties in LLMs via Beam Search. In *The Fourteenth International Conference on Learning Representations*.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. [Detecting hallucinations in large language models using semantic entropy](#). *Nature*, 630(8017):625–630.
- Constanza Fierro and Anders Søgaard. 2022. [Factual consistency of multilingual pretrained language models](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.
- Lukas Fluri, Daniel Paleka, and Florian Tramèr. 2024. [Evaluating Superhuman Models with Consistency Checks](#). In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 194–232.
- David C. Funder and C. Randall Colvin. 1991. [Explorations in behavioral consistency: Properties of persons, situations, and behaviors](#). *Journal of Personality and Social Psychology*, 60(5):773–794.
- Prakhar Ganesh, Reza Shokri, and Golnoosh Farnadi. 2026. [Rethinking Hallucinations: Correctness, Consistency, and Prompt Multiplicity](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6959–6978, Rabat, Morocco. Association for Computational Linguistics.
- Bishwamitra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. 2025. Logical Consistency of Large Language Models in Fact-Checking. In *The Thirteenth International Conference on Learning Representations*.
- Yuling Gu, Bhavana Dalvi Mishra, and Peter Clark. 2023. [Do language models have coherent mental models of everyday things?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1892–1913, Toronto, Canada. Association for Computational Linguistics.
- Lovisa Hagström, Denitsa Saynova, Tobias Norlund, Moa Johansson, and Richard Johansson. 2023. [The effect of scaling, retrieval augmentation and form on the factual consistency of language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5457–5476, Singapore. Association for Computational Linguistics.
- Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. [Methods for measuring, updating, and visualizing factual beliefs in language models](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2714–2731, Dubrovnik, Croatia. Association for Computational Linguistics.
- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. [TRUE: Re-evaluating factual consistency evaluation](#). In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.
- Shomik Jain, D. Calacci, and Ashia Wilson. 2024. [As an AI Language Model, "Yes I Would Recommend](#)

- Calling the Police": Norm Inconsistency in LLM Decision-Making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):624–633.
- Dongjun Jang, Youngchae Ahn, and Hyopil Shin. 2025. **RCScore: Quantifying Response Consistency in Large Language Models**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5690–5708, Suzhou, China. Association for Computational Linguistics.
- Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. **BECEL: Benchmark for consistency evaluation of language models**. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Myeongjun Jang and Thomas Lukasiewicz. 2023. **Consistency analysis of ChatGPT**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15970–15985, Singapore. Association for Computational Linguistics.
- Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo Jose Taylor, and Dan Roth. 2024. **A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4722–4756, Miami, Florida, USA. Association for Computational Linguistics.
- Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Yuan Yuan, Zhuoqun Hao, Xinyi Bai, Weijie J Su, Camillo Jose Taylor, and Tanwi Mallick. 2025. **Towards Rationality in Language and Multimodal Agents: A Survey**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3656–3675, Albuquerque, New Mexico. Association for Computational Linguistics.
- Hongjian Jiang, Michael Hahn, Georg Zetsche, and Anthony Widjaja Lin. 2026. **Softmax Transformers are Turing-Complete**. In *The Fourteenth International Conference on Learning Representations*.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. **How can we know what language models know?** *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. **When to make exceptions: Exploring language models as accounts of human moral judgment**. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Jana Jung, Marlene Lutz, Indira Sen, and Markus Strohmaier. 2026. **Do Psychometric Tests Work for Large Language Models? Evaluation of Tests on Sexism, Racism, and Morality**. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8143–8173, Rabat, Morocco. Association for Computational Linguistics.
- Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. **Why Language Models Hallucinate**.
- Nora Kassner and Hinrich Schütze. 2020. **Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.
- Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. 2023. **Language models with rationality**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14190–14201, Singapore. Association for Computational Linguistics.
- Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. **BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8849–8861, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Niklas Kiehne, Hermann Kroll, and Wolf-Tilo Balke. 2022. **Contextualizing language models for norms diverging from social majority**. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4620–4633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Niklas Kiehne, Alexander Ljapunov, Marc Bätje, and Wolf-Tilo Balke. 2024. **Analyzing effects of learning downstream tasks on moral bias in large language models**. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 904–923, Torino, Italia. ELRA and ICCL.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. **Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation**. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiušė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver

- Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. [Measuring Faithfulness in Chain-of-Thought Reasoning](#).
- Noah Lee, Jiwoo Hong, and James Thorne. 2025. Evaluating the Consistency of LLM Evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10650–10659, Abu Dhabi, UAE. Association for Computational Linguistics.
- Steven A. Lehr, Ketan S. Saichandran, Eddie Harmon-Jones, Nykko Vitali, and Mahzarin R. Banaji. 2025. [Kernels of selfhood: GPT-4o shows humanlike patterns of cognitive dissonance moderated by free choice](#). *Proceedings of the National Academy of Sciences*, 122(20):e2501823122.
- Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikrumar. 2019. [A logic-driven framework for consistency of neural models](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.
- Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024a. [Can multiple-choice questions really be useful in detecting the abilities of LLMs?](#) In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia. ELRA and ICCL.
- Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. 2024b. [Benchmarking and improving generator-validator consistency of language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, and Zhiyu Li. 2024. [Internal Consistency and Self-Feedback in Large Language Models: A Survey](#).
- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research*.
- Andy Liu, Kshitish Ghate, Mona T. Diab, Daniel Fried, Atoosa Kasirzadeh, and Max Kleiman-Weiner. 2026. Generative Value Conflicts Reveal LLM Priorities. In *The Fourteenth International Conference on Learning Representations*.
- Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. 2025. Aligning with Logic: Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models. In *Forty-Second International Conference on Machine Learning*.
- Lea Löhn, Niklas Kiehne, Alexander Ljapunov, and Wolf-Tilo Balke. 2024. [Is Machine Psychology here? On Requirements for Using Human Psychological Tests on Large Language Models](#). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 230–242, Tokyo, Japan. Association for Computational Linguistics.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. [SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.
- William Merrill and Ashish Sabharwal. 2023. [The parallelism tradeoff: Limitations of log-precision transformers](#). *Transactions of the Association for Computational Linguistics*, 11:531–545.
- William Merrill and Ashish Sabharwal. 2024. [The expressive power of transformers with chain of thought](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Marcus J. Min, Yangruibo Ding, Luca Buratti, Saurabh Pujar, Gail E. Kaiser, Suman Jana, and Baishakhi Ray. 2024. [Beyond accuracy: Evaluating self-consistency of code large language models with identitychain](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher Manning. 2022. [Enhancing self-consistency and performance of pre-trained language models through natural language inference](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1768, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. [State of What Art? A Call for Multi-Prompt LLM Evaluation](#). *Transactions of the Association for Computational Linguistics*, 12:933–949.
- Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. [Are Large Language Models Consistent over Value-laden Questions?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221, Miami, Florida, USA. Association for Computational Linguistics.
- Benjamin Newman, Prafulla Kumar Choubey, and Nazneen Rajani. 2022. [P-adapters: Robustly extracting factual information from language models with diverse prompts](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

- Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. [Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Jekaterina Novikova, Carol Myrick Anderson, Borhane Blili-Hamelin, Domenic Rosati, and Subhabrata Majumdar. 2025. Consistency in Language Models: Current Landscape, Challenges, and Future Directions. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.
- José Luiz Nunes, Guilherme F. C. F. Almeida, Marcelo de Araujo, and Simone D. J. Barbosa. 2024. [Are Large Language Models Moral Hypocrites? A Study Based on Moral Foundations](#). *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1074–1087.
- Maxwell I. Nye, Michael Henry Tessler, Joshua B. Tenenbaum, and Brenden M. Lake. 2021. [Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning](#). In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 25192–25204.
- Xenia Ohmer, Elia Bruni, and Dieuwke Hupke. 2024. [From Form\(s\) to Meaning: Probing the Semantic Depths of Language Models Using Multisense Consistency](#). *Computational Linguistics*, 50(4):1507–1556.
- Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2023. [Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 258–276, Singapore. Association for Computational Linguistics.
- Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. 2025. Consistency Checks for Language Model Forecasters. In *The Thirteenth International Conference on Learning Representations*.
- Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. [Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.
- Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. [Cross-lingual consistency of factual knowledge in multilingual language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.
- Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby Tavor. 2023. [Predicting question-answering performance of large language models through semantic consistency](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 138–154, Singapore. Association for Computational Linguistics.
- Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. 2024. Improving Consistency in Large Language Models through Chain of Guidance. *Transactions on Machine Learning Research*.
- Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2022. Measuring Reliability of Large Language Models through Semantic Consistency. In *NeurIPS ML Safety Workshop*.
- Sai Sathiesh Rajan, Ezekiel Soremekun, and Sudipta Chattopadhyay. 2024. [Knowledge-based Consistency Testing of Large Language Models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10185–10196, Miami, Florida, USA. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. [Are red roses red? evaluating consistency of question-answering models](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.
- Jillian Ross, Yoon Kim, and Andrew Lo. 2024. LLM economicus? Mapping the Behavioral Biases of LLMs via Utility Theory. In *First Conference on Language Modeling*.
- Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. [Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.
- Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. [Evaluating the moral beliefs encoded in llms](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero,

- Marwa Abdulhai, Aleksandra Faust, and Maja Mataric. 2025. [A psychometric framework for evaluating and shaping personality traits in large language models](#). *Nature Machine Intelligence*, 7(12):1954–1968.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models](#).
- Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. [Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118, Suzhou, China. Association for Computational Linguistics.
- Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. [You don’t need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. [Position: A roadmap to pluralistic alignment](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Jakob Stenseke, Aidan Kierans, Itamar Pres, and Dylan Hadfield-Menell. 2026. [Position: Why LLMs should be reasonably morally inconsistent](#). In *Pluralistic Alignment Workshop at ICML 2026*.
- Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. 2025. [Too Consistent to Detect: A Study of Self-Consistent Errors in LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4755–4765, Suzhou, China. Association for Computational Linguistics.
- Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025. CoCoA: A Minimum Bayes Risk Framework Bridging Confidence and Consistency for Uncertainty Quantification in LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. [Large Language Models are not Fair Evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Rottger, and Barbara Plank. 2024b. Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think. In *First Conference on Language Modeling*.
- Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024c. [“My Answer is C”: First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.
- Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. [Measure and improve robustness in NLP models: A survey](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Xiaoyuan Wu, Weiran Lin, Omer Akgul, and Lujia Bauer. 2025. [Estimating LLM Consistency: A User Baseline vs Surrogate Metrics](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30530–30544, Suzhou, China. Association for Computational Linguistics.
- Zhihui Xie, Jizhou Guo, Tong Yu, and Shuai Li. 2024. [Calibrating reasoning in language models with internal consistency](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. [Knowledge Conflicts for LLMs: A Survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.
- Ruoxi Xu, Hongyu Lin, Xianpei Han, Jia Zheng, Weixiang Zhou, Le Sun, and Yingfei Sun. 2025. Large Language Models Often Say One Thing and Do Another. In *The Thirteenth International Conference on Learning Representations*.

- Jingyuan Yang, Dapeng Chen, Yajing Sun, Rongjun Li, Zhiyong Feng, and Wei Peng. 2024a. [Enhancing Semantic Consistency of Large Language Models through Model Editing: An Interpretability-Oriented Approach](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3343–3353, Bangkok, Thailand. Association for Computational Linguistics.
- Zhe Yang, Yichang Zhang, Tianyu Liu, Jian Yang, Junyang Lin, Chang Zhou, and Zhifang Sui. 2024b. [Can Large Language Models Always Solve Easy Problems if They Can Solve Harder Ones?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1531–1555, Miami, Florida, USA. Association for Computational Linguistics.
- Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. [Benchmarking llms via uncertainty quantification](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Haoran Ye, TianZe Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, and Guojie Song. 2025. [Generative Psycho-Lexical Approach for Constructing Value Systems in Large Language Models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11968–11991, Vienna, Austria. Association for Computational Linguistics.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? In *2nd AI for Math Workshop @ ICML 2025*.
- Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. [R-tuning: Instructing large language models to say ‘I don’t know’](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.
- Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024b. [How language model hallucinations can snowball](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Shuaiqiang Wang, Chong Meng, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. [Improving the robustness of large language models via consistency alignment](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8931–8941, Torino, Italia. ELRA and ICCL.
- Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. [Prompt consistency for zero-shot task generalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2613–2626, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Di Zhou and Yinxian Zhang. 2024. [Political biases and inconsistencies in bilingual GPT models—the cases of the U.S. and China](#). *Scientific Reports*, 14(1):25048.
- Ziyi Zhou, Xinwei Guo, Jiashi Gao, Xiangyu Zhao, Shiyao Zhang, Xin Yao, and Xuetao Wei. 2024. [Unveiling the bias impact on symmetric moral consistency of large language models](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

A Further Coding Details

For our analysis of consistency evaluations in Table 2, we relied on a strict coding scheme with binary decisions. We conceptualized the coding process as deliberately permissive, assigning ✓ if the methodology featured any of our three requirements, regardless of the depth of the evaluation. Additionally, if a paper featured multiple evaluations, we prioritized the assignment of ✓ over ✗. The following provides detailed descriptions for each requirement:

- R1:** We use ✗ to indicate tasks with clearly preferred/correct answers. This includes any factual knowledge, QA or reasoning task, LLM-as-a-judge tasks with labeled preferences, and also any data for which the authors or dataset creators provide or test against an a priori ground truth (note that this does not have to mean that the authors actually evaluate any correctness in their paper). ✓ signifies tasks with no a priori clearly correct answers.
- R2:** We use ✗ to indicate direct, point-wise output comparisons that do not account for variance. ✓ signifies measurements comparing aggregations of multiple samples or token distributions between inputs (e.g., divergence metrics, comparison of answer distributions or token rankings). This does not require the evaluation to actually quantify any uncertainty.
- R3:** We use ✗ to indicate an evaluation only testing responses that are forced commitments in the first word or token. ✓ signifies additional evaluation that allows for either sequential reasoning/CoT before answering or open-ended responses.

B Detailed Literature Overview for Semantic and Relational Consistency

Below, we provide the detailed itemization of the semantic and relational consistency evaluations parallel to the sampling consistency overview provided in Table 1 of the main text.

Table 3 outlines the 51 publications evaluating semantic consistency, as discussed in Section 3.2.

Table 4 details the 40 publications evaluating relational consistency, categorized into logical, conceptual, and psychometric consistency constraints, as defined in Section 3.3.

The columns are coded as follows:

CONSISTENCY CHECKS: **Bold** text indicates a task in ambiguous or subjective domains without a priori correct answers, fulfilling **R1**.

METRICS: **Bold** text indicates a metric that disentangles uncertainty using probabilistic comparisons rather than single-sample matches, fulfilling **R2**.

OUTPUT: Direct answers (i.e., any commitment in the first token) are in normal text, ***bold italics*** for chain-of-thought (CoT), and **bold** for open-ended outputs requiring heuristic extraction, annotated by extraction method, e.g., man (manual), NLI (natural language inference models), emb (embedding similarity), CE (cross-encoder similarity), LLM (LLM extraction). Multiple heuristics are marked with *, with further details about the methods provided in the **METRIC** column.

PAPER (<i>Venue</i>)	CONSISTENCY CHECKS	METRIC	OUTPUT
Li et al. (2019) (<i>EMNLP-IJCNLP</i>)	Symmetric contradiction in NLI tasks	% of exmpl. w/ constraint violation (Global Violation) + where antecedent holds (Conditional Violation)	BERT NLI
Asai and Hajishirzi (2020) (<i>ACL</i>)	Compl. answer for antonyms or negated questions	Glob. Viol. (See Li et al., 2019)	BERT QA
Jiang et al. (2020) (<i>TACL</i>)	Same entity for paraphr. expres. the same KB relation	% of cons. inputs for 2 prompts	Single-token
Kassner and Schütze (2020) (<i>ACL</i>)	Cons. toward negations and misprimes in KB relat.	Neg.: Spearman rank corr. ; Mispr.: Drop in prec.	Token ranking
Elazar et al. (2021) (<i>TACL</i>)	Same entity for paraphr. expres. the same KB relation	Avg. % of cons. prompt pairs & % where all ans. correct	Single-token
Newman et al. (2022) (<i>ICLR</i>)	Same entity for paraphr. expres. the same KB relation	See (Elazar et al., 2021)	Single-token
Fierro and Søgaard (2022) (<i>ACL Findings</i>)	Same entity for paraphr. expres. the same multiling. KB relat.	See (Elazar et al., 2021)	Single-token
Honovich et al. (2022) (<i>NAACL</i>)	Factual cons. btw. open-ended texts (summarization, fact verification, paraphrases, dialogue)	ROUGE/BLEU, BERTScore, NLI, QG-QA	Open-ended*
Jang et al. (2022) (<i>COLING</i>)	Cons. answer for paraphr., order var. and negations	% of cons. input pairs	True/False classif.
Albrecht et al. (2022) (<i>NeurIPS MLSW</i>)	Reworded commons. moral quest.	% of cons. reworded inputs	(Not) Wrong
Raj et al. (2022) (<i>NeurIPS MLSW</i>)	Sim. open-ended answ. to paraphr. questions	Avg. BERTScore, paraphr. pred. score, NLI-score	Open-ended*
Kiehne et al. (2022) (<i>EMNLP Findings</i>)	Action choices cons. w/ contrary social norms after fine-tuning	Accuracy diff. after fine-tuning	Binary classif.
Zhou et al. (2022) (<i>EMNLP Findings</i>)	Same classif. under paraphrased input and labels	Fleiss' kappa over paraphrases and inputs	MCQ
Hase et al. (2023) (<i>EACL</i>)	Paraphrase cons. in QA	See (Elazar et al., 2021)	True/False
Scherrer et al. (2023) (<i>NeurIPS</i>)	Same action wrt. prompt rephrasing & option order var. in moral scen.	JSD btw. answer distribution and average	Binary-choice
Hagström et al. (2023) (<i>EMNLP</i>)	1. Same entity for paraphr. expres. the same KB relat. 2. Paraphr. cons. of retrieval	1. See (Elazar et al., 2021); 2. Overlap of retriev. passages; cos. sim. of retriev. emb.	Single-word Retr. pass. & embed.
Jang and Lukasiewicz (2023) (<i>EMNLP</i>)	Cons. answer for paraphr., order var. and negations	See (Jang et al., 2022)	Yes/No
Qi et al. (2023) (<i>EMNLP</i>)	Same token rankings for prompts in 2 languages expres. same KB rel.	RankC: Avg. weighted sum of intersec. sizes of the top j tokens for both prompts	Token ranking
Ohmer et al. (2023) (<i>GEM</i>)	Same answer for self-transl. input	% of cons. pairs rel. to sampl.	Yes/No
Rabinovich et al. (2023) (<i>GEM</i>)	Sim. factoid answ. to paraphr. quest.	Avg. cos. sim. btw. emb. answer pairs	Single-word/phrase ^{emb}
Fluri et al. (2024) (<i>SaTML</i>)	1. Paraphr. & negation cons. of forecasting probs. ; 2. BERT classf. on paraphr. in human rights violations	1. (Inverted) diff. of probs.; 2. Change in class probs. and % cases w/ switched classes	1. CoT + Prob. ; 2. Binary classif.
Agarwal et al. (2024) (<i>LREC-COLING</i>)	Cons. action agreement in moral dilem. in diff. languages	Comp. of Monte-Carlo answer distrib. btw. languages	MCQ
Bonagiri et al. (2024) (<i>LREC-COLING</i>)	Sim. open-ended answers for paraphr. moral questions	SaGE: Graph entropy of weighted cos. sim. of answers	Open-ended^{emb}
Kiehne et al. (2024) (<i>LREC-COLING</i>)	Cons. wrt. prompt var. of moral norm judgem.	Avg., max., min., std. dev. of % of cons. prompt pairs	Single-token
Li et al. (2024a) (<i>LREC-COLING</i>)	1. MCQ order cons.; 2. Cons. btw. MCQ & open-ended text answ.	1. Chi-sq. statistic & accur. diff.; 2. Conf. matrix & Pearson's corr.	MCQ; open-ended^{man}
Zhao et al. (2024) (<i>LREC-COLING</i>)	Cons. open-ended answ. for paraphr. instruct.	Avg. % of cons. answ. pairs & of maj. answ. rated by GPT-4	Open-ended^{LLM}
Shu et al. (2024) (<i>NAACL</i>)	Symbol, format changes & option, negation, order cons. in person. tests	% same response	Agree/Disagree

PAPER (<i>Venue</i>)	CONSISTENCY CHECKS	METRIC	OUTPUT
Ceron et al. (2024) (<i>TACL</i>)	Paraphr./negation/order/templ. cons. on political opinion quest.	% same response & Krippendorff's α	Agree/Disagree
Röttger et al. (2024) (<i>ACL</i>)	1. 10 paraphr. instructions; 2. MCQ vs open-ended text for pol. compass quest.	1. % of cons. answ. & change of pol. compass position; 2. Answ. distr. change	MCQ; open-ended ^{LLM}
Wang et al. (2024a) (<i>ACL</i>)	Option order cons. of binary LLM as judge preferences	% conflicting judgem.	Pref. score; CoT + score
Wang et al. (2024c) (<i>ACL Findings</i>)	1. First-token MCQ vs open-ended text; 2. Order cons. (MMLU + OpinionQA)	1. % same response; 2. Entropy over 10 random orders	MCQ; open-ended ^{LLM}
Yang et al. (2024a) (<i>ACL Findings</i>)	Paraphr. cons. on NLU and NLG tasks	NLU: Std. dev. of acc.; NLG: Answ. sim. rated by GPT-4	MCQ; open-ended ^{LLM}
Mizrahi et al. (2024) (<i>TACL</i>)	Robustn. of perform. rank against instruction paraphr.	1. Kendall's W; 2. Kendall's τ ; 3. Std. dev. from avg. prompt perform.	Single-word/phrase
Durmus et al. (2024) (<i>COLM</i>)	Changing opinion agreem. w/ country when prompted in other language	Change in JSD btw. LLM and population answer distr.	MCQ token probs.
Wang et al. (2024b) (<i>COLM</i>)	1. Random perturb. cons.; 2. First-token MCQ vs open-ended text; 3. Change w/ addit. neutral options	1. Entropy over all random perturb.; 2. & 3. % same response	MCQ; open-ended ^{LLM}
Zhou and Zhang (2024) (<i>Nature Scientific Reports</i>)	Cons. content and sentiment btw. English and Chinese political quest.	% cons. response; Diff. in sentim. score btw. languages	Yes/No; open-ended ^{man}
Raj et al. (2024) (<i>TMLR</i>)	Sim. open-ended answ. to paraphr. questions	Avg. NLI entailm., paraphr. pred. score, ROUGE-L	Open-ended*
Moore et al. (2024) (<i>EMNLP Findings</i>)	Paraphr., multiling., MCQ vs open-ended text on value quest.	JSD btw. token distrib. and centroid (binary for open-ended)	Yes/No tkn. probs.; open-ended ^{LLM}
Rajan et al. (2024) (<i>EMNLP Findings</i>)	Same answ. for paraphr. KG fact	% same response	Yes/No
Zhou et al. (2024) (<i>NeurIPS</i>)	Same action wrt. option letter & order swap in moral scen.	% same response & t-SMC	Single-tkn. binary-ch.
Ohmer et al. (2024) (<i>CL</i>)	Same answ. for self-paraphr./-transl. input	% of cons. pairs rel. to sampl.	Yes/No, Single-word
Lee et al. (2025) (<i>COLING</i>)	Cons. scores in text evaluation btw. diff. eval. scales	Krippendorff's α btw. scale types	Num. & Likert sc., Y/N.
Liu et al. (2025) (<i>ICML</i>)	Commutativity and negation cons. of LLM as judge preferences	% of cons. pairs	Binary-ch. (+ CoT)
Calanzone et al. (2025) (<i>ICLR</i>)	Complem. answ. for negation of KG fact quest.	Cond. Viol. (See Li et al., 2019)	True/False; CoT + T/F.
Ghosh et al. (2025) (<i>ICLR</i>)	Complem. answ. for negation of KG fact quest.	% complementary response	True/False; CoT + T/F.
Paleka et al. (2025) (<i>ICLR</i>)	Cons. forecasting probs. wrt. paraphr. and negated scenarios	Arbitrage & frequentist probabilistic cons. violation	Score only; CoT + Score
Jang et al. (2025) (<i>EMNLP</i>)	Consistent open-ended responses for stylistic prompt variations in math QA	Score based on structural, lexical and logical overlap of 2 respon.	Open-ended*
Bu et al. (2026) (<i>ICLR</i>)	Consistently safe or unsafe answers across mult. lang.	Like (Elazar et al., 2021) for same safety rating	Open-ended ^{LLM}
Liu et al. (2026) (<i>ICLR</i>)	Same value pref. rankings btw. MCQ and open-ended generation answers	Difference in preference rankings of values	MCQ; Open-ended ^{LLM}
Ganesh et al. (2026) (<i>EACL</i>)	Same response for paraphrases, example order & MCQ option order changes	1. % quest. w/ >0 different resp.; 2. % of diff. response pairs	MCQ
Jung et al. (2026) (<i>EACL</i>)	Rephras., option order, symbol cons. in psychom. racism/sexism/moral. tests	% same response	Likert scales

Table 3: Overview of methods in semantic consistency evaluations. **Bold** text in the **CONSISTENCY CHECKS** column indicates tasks without a ground truth (**R1**). **Bold** text in the **METRIC** column indicates distributional comparisons between outputs (**R2**). **Bold italic** text in the **OUTPUT** column represents outputs with CoT, **bold** text represents open-ended generation with the respective extraction heuristic provided in superscript (**R3**).

PAPER (<i>Venue</i>)	CONSISTENCY CHECKS	METRIC	OUTPUT
Logical Consistency			
Li et al. (2019) (<i>EMNLP-IJCNLP</i>)	Transitivity of implications in NLI tasks	% of exmpl. w/ constraint violation (Global Violation) + where antecedent holds (Conditional Violation)	BERT NLI
Ribeiro et al. (2019) (<i>ACL</i>)	Robustn. to validation quest. entailed by QA pair	% of answ. cons. w/ original ans.	Single-word; Yes/No
Asai and Hajishirzi (2020) (<i>ACL</i>)	Transitivity of cause-effect QA	Glob. Violat. (Li et al., 2019)	BERT QA
Kassner et al. (2021) (<i>EMNLP</i>)	Impl. cons. of binary KG fact quest.	Condit. Violat. (Li et al., 2019)	True/False
Nye et al. (2021) (<i>NeurIPS</i>)	Neuro-symbol. cons. of step-by-step stories or relationships	% of lines & full outputs w/out constraint violation	Fixed-format text
Jang et al. (2022) (<i>COLING</i>)	Impl. transit. in cause-effect QA and NLI; Conj. of quest. w/ same label	Condit. Violat. (Li et al., 2019)	True/False classif.
Mitchell et al. (2022) (<i>EMNLP</i>)	Impl. cons. of binary KG fact quest.	Condit. Violat. (Li et al., 2019)	True/False
Hase et al. (2023) (<i>EACL</i>)	1. Impl. & contrapos. cons. on binary quest.; 2. Transit. belief change after knowl. editing	1. Condit. Violat. (Li et al., 2019); 2. % transitive triples	True/False classif.
Gu et al. (2023) (<i>ACL</i>)	Cons. (a-)symmetric, inverse, transitive spatial relat. in physical objects	Condit. Violat. (Li et al., 2019)	True/False
Jang and Lukasiewicz (2023) (<i>EMNLP</i>)	Transitivity of impl. in NLI and cause-effect QA	Condit. Violat. (Li et al., 2019)	Yes/No
Kassner et al. (2023) (<i>EMNLP</i>)	Cons. beliefs btw. answ. options of MCQ quest.	Condit. Violat. (Li et al., 2019)	True/False
Fluri et al. (2024) (<i>SaTML</i>)	1. Forecasting probs. follow monotonicity and Bayes Rule; 2. Impl. cons. on legal bail decisions	1. Spearman rank corr. (for monot.); Bayes Rule; 2. Cond. Violat. (Li et al., 2019)	CoT + Prob./Yes/No
Ross et al. (2024) (<i>COLM</i>)	1. Monotonicity in economic decisions ; 2. Cons. better/worse econ. rationality than humans	1. % cons. responses; 2. Compar. to human popul. over mult. experiments	Single-word; CoT + Single-word
Rajan et al. (2024) (<i>EMNLP</i>)	Transitive cons. of KG fact quest.	% cons. responses	Yes/No
Calanzone et al. (2025) (<i>ICLR</i>)	(Reverse) impl. cons. of KG fact quest.	Condit. Violat. (Li et al., 2019)	True/False; CoT + T/F.
Ghosh et al. (2025) (<i>ICLR</i>)	Disj., conj. & complex logic rule cons. of KG fact quest.	% inputs w/ cons. answ.	True/False; CoT + T/F.
Paleka et al. (2025) (<i>ICLR</i>)	Cons. forecasting probs. wrt. different prob. constraints	Arbitrage & frequentist probabilistic cons. violation	Score only; CoT + Score
Liu et al. (2025) (<i>ICML</i>)	Transitivity cons. of LLM as judge preferences	Ratio of subgraphs of the pref. graph that are acyclic	(CoT) Bin.-choice
Conceptual Consistency			
Jin et al. (2022) (<i>NeurIPS</i>)	Conservativeness in making exceptions on moral judgem.	% of judgem. agreeing w/ underlying principle	Yes/No
Chen et al. (2023) (<i>TMLR</i>)	1. Same answ. for quest. and hypothet. quest.; 2. Cons. answ. btw. partial quest. and composite of quest.	1. % same answ.; 2. % compos. quest. reusing the partial answ.	MCQ
Berglund et al. (2024) (<i>ICLR</i>)	Symmetric entity knowl. when asked for relation	Diff. of acc. on relation and reverse relation	Single-word/-token
Li et al. (2024b) (<i>ICLR</i>)	Cons. of LLM when generating and validating own answ.	% cons. responses	Fixed-format text
Min et al. (2024) (<i>ICLR</i>)	Cons. code after n translations to natural language and back	% test cases w/ matching error message	Code; Open-ended
Zhang et al. (2024b) (<i>ICML</i>)	Cons. btw. answer and the LLM validating its rationale	% rationales the LLM validates as wrong	Yes/No; CoT + Yes/No
Jain et al. (2024) (<i>AIES</i>)	Cons. btw. factual & normative quest. about surveillance footage	% of Yes/No/ambig. answers in both settings	Yes/No + explanation

PAPER (<i>Venue</i>)	CONSISTENCY CHECKS	METRIC	OUTPUT
Jiang et al. (2024) (<i>EMNLP</i>)	Cons. answ. on fallacy quest. against logically irrelevant changes	% correct before & after changes	<i>(CoT)</i> Bin.-choice
Moore et al. (2024) (<i>EMNLP</i>)	Cons. answ. btw. value quest. on the same topic	JSD btw. answ. distr. and centroid	Yes/No token probs.
Yang et al. (2024b) (<i>EMNLP</i>)	Correct answ. for easy quest. if LLM correctly answ. harder one	Cond. prob. of answ. easy q. given prob. for hard q.	Fixed-format text
Xu et al. (2025) (<i>ICLR</i>)	Cons. btw. abstract value & concrete scenario judgem.	% cons. answ.; JSD btw. first token distr.	A/B; <i>CoT + A/B</i>
Bajpai and Chakraborty (2025) (<i>EMNLP</i>)	Same answer for event-oriented and time-oriented historical questions	% cons. pairs	Entity name; <i>CoT + ent. n.</i>
Ding et al. (2025) (<i>EMNLP</i>)	Cons. detection of harmful prompts both when tasked to generate and discriminate	% pairs where LLM detects harm in one but not the other	Open-ended ^{LLM}
Shen et al. (2025) (<i>EMNLP</i>)	Action choices cons. with value agreem.	F1 of cons. answ.; Dist. btw. mean action and value choice	Likert scales & MCQ
Psychometric Consistency			
Santurkar et al. (2023) (<i>ICML</i>)	Cons. agreem. w/ one group on public opinion quest.	% of topics where token distrib. agrees w/ the most aligned group	MCQ token distrib.
Ceron et al. (2024) (<i>TACL</i>)	Cons. political stance on a topic	% agree - % disagree per topic	Agree/Disagree
Durmus et al. (2024) (<i>COLM</i>)	Change of opinion distrib. agreem. w/ a country when prompted to answer like a person from that country	Change in JSD btw. LLM and population answers	MCQ token probs.
Nunes et al. (2024) (<i>AIES</i>)	1. Internal cons. within moral questionnaire ; 2. Correl. between value and vignette quest.	1. Cronbach’s α ; 2. Ordinary linear regression (over 100 tests)	Full questionnaires w/ Likert sc.
Lehr et al. (2025) (<i>PNAS</i>)	1. Cons. btw. pol. opinion quest. and sentiment in open-ended text essay; 2. Internal cons./reliability	1. Deviation from mean z-score ; 2. Cronbach’s α	Likert scales
Ye et al. (2025) (<i>ACL</i>)	1. Internal cons. within value questionnaire ; 2. Correl. of LLM values btw. topics and two datasets	1. Cronbach’s α ; 2. Correl. of extracted values	Open-ended ^{LLM}
Serapio-García et al. (2025) (<i>Nature Machine Intelligence</i>)	Reliability/internal cons. on psychometric tests	Cronbach’s α , Guttman’s λ_6 , McDonald’s ω	Likert scales
Jung et al. (2026) (<i>EACL</i>)	1. Correl. scores btw. related psychom. tests ; 2. Test scores correlate with open-ended text answ. in related downstr. tasks	Spearman’s rank correlation	Likert scales; open-ended*

Table 4: Overview of methods in relational consistency evaluations. **Bold** text in the **CONSISTENCY CHECKS** column indicates tasks without a ground truth (**R1**). **Bold** text in the **METRIC** column indicates distributional comparisons between outputs (**R2**). **Bold italic** text in the **OUTPUT** column represents outputs with CoT, **bold** text represents open-ended generation with the respective extraction heuristic provided in superscript (**R3**). The literature is categorized into logical, conceptual, and psychometric consistency, as defined in Section 3.3.