Benchmarking LLM Overshooting: Automatic Evaluation of Refutation Quality and Emotional Alignment under Pressure

Anonymous authors
Paper under double-blind review

Abstract

Despite the substantial convenience and productivity gains provided by large language models (LLMs), a growing concern is users' uncritical or blind trust in their generated responses. Such emotional alignment has recently attracted attention, with a specific focus on overshooting—a phenomenon in which users attribute emotional value to artificial intelligence beyond its inherent capabilities. Despite recent advancements, "refutation quality" and "emotional alignment" remain largely unquantified in situations where LLMs encounter false premises. To address this gap, we introduce a new benchmark that enables automatic quantification of LLM overshooting. Specifically, it defines an Overshoot Index (OI) that integrates six metrics: Refutation Strength (RS), Directness Index (DI), Hedging Load (HL), Affective Overshoot Proxy (AOP), Normative Jump (NJ), and Evidence-Backed Correction (EBC). In our experiments, three models—OpenAI's `gpt-4o-mini`, Anthropic's `claude-3-5-sonnet-20241022`, and Google's `gemini-1.5-flash`—were evaluated using prompts generated from the TruthfulQA, CREPE, and FalseQA datasets. Additionally, three pressure levels (pressure $\in \{0, 1, 2\}$) were introduced to examine behavioral changes under stress. Rather than ranking models, OI serves as a **diagnostic benchmark** that reveals how refutation strength and emotional accommodation interact under false-premise conditions. Across the three commercial LLMs, OI highlights distinct behavioral tendencies, illustrating its value as a complementary tool for alignment and safety evaluation rather than a performance leaderboard. Statistical validation was performed using Kruskal–Wallis and Wilcoxon signed-rank tests. Overall, this study provides a novel perspective for evaluating LLM safety and robustness.

1 Introduction

Large language models (LLMs) are increasingly being applied for search assistance, text summarization, and tasks involving human decision-making and emotions, such as education, healthcare, legal services, and counseling. Although this societal adoption offers significant convenience and productivity gains, it also introduces the risk of users placing excessive trust in or developing emotional dependence on generated responses. Understanding how LLMs respond to questions containing false premises is a crucial issue that extends beyond simple correctness determination. Such responses risk leading users to believe in misinformation and may encourage inappropriate emotional investments in their interactions with artificial intelligence (AI).

Existing research primarily addresses factuality, hallucination suppression, and harmful query refusal. Benchmarks such as TruthfulQA are widely used to quantify the correctness of generated responses (Lin et al., 2022), while frameworks such as HELM and BIG-Bench evaluate model utility and safety comprehensively (Liang et al., 2022; Srivastava et al., 2022). Additionally, model-tuning strategies such as Anthropic's Constitutional AI have been developed to enhance safety and mitigate harm (Bai et al., 2022). However, these evaluations do not explicitly measure "refutation quality" or "emotional alignment." The following question, therefore, remains unanswered: How does an LLM refute, and how does it respond emotionally when confronted with false premises?

The concept of emotional alignment has recently gained attention in philosophical and ethical discussions. Schwitzgebel and Sebo (Schwitzgebel & Sebo, 2025) proposed an "Emotional Alignment Design Policy," arguing that AI-elicited emotional responses in humans should correspond to the system's actual capabilities and moral status. Within this context, *overshooting* is particularly problematic. This phenomenon occurs when users attribute emotional value to AI beyond its inherent worth, potentially leading to excessive attachment, overconfidence, or economic investment. Recent studies highlight the risk that strong emotional bonding with AI chatbots may foster psychological dependence and social isolation, posing a broader social concern.

To address these gaps, this paper introduces a benchmark for automatically quantifying LLM overshooting. The *Overshoot Index (OI)* is defined by integrating six submetrics: Refutation Strength (RS), Directness Index (DI), Hedging Load (HL), Affective Overshoot Proxy (AOP), Normative Jump (NJ), and Evidence-Backed Correction (EBC). This framework simultaneously evaluates both the negation strength and the degree of affective accommodation. This methodology combines automatic detection based on regular expressions and vocabulary lists, inference-based implication determination via Hugging Face's natural language inference (NLI) models, and verification of external knowledge consistency through the Wikipedia API, thereby eliminating the need for human annotation.

To the best of the author's knowledge, OI is the first benchmark to disentangle refutation quality and emotional overshooting in a fully automated and reproducible manner. This enables systematic and transparent evaluation that was not achievable with prior approaches focusing on a single dimension.

The key contributions of this study are as follows: (i) the automatic separation and quantification of "quality of negation" and "emotional compliance," which existing benchmarks cannot capture; (ii) the simulation of realistic scenarios through the integration of multiple data sources and pressure conditions; and (iii) the establishment of statistically significant differences, facilitating reliable discussion of model characteristics.

The remainder of this paper is organized as follows: Section 2 reviews related research, Section 3 describes the proposed method, Section 4 presents the experimental results and statistical validation, and Section 6 concludes the paper and discusses future challenges.

This work does not aim to rank commercial systems but to introduce a reproducible, interpretable, and diagnostic benchmark that helps researchers understand how LLMs balance refutation strength and affective alignment. OI is therefore intended as a **complementary evaluation tool** alongside factuality, hallucination, and refusal benchmarks, highlighting underexplored behavioral dimensions rather than competitive performance.

2 Related Work

2.1 Factuality and Hallucination Evaluation

Extensive research has examined factual consistency and hallucination detection in LLMs. Benchmarks such as WikiContradict and DefAn capture real-world knowledge conflicts and provide definitive-answer settings for hallucination evaluation (Hou et al., 2024; Rahman et al., 2024). However, several studies have highlighted the limitations of existing hallucination detection methods, noting that many metrics either fail to correlate with human judgments or focus too narrowly on lexical overlap (Janiak et al., 2025; Luo et al., 2024). Recent work further emphasizes the importance of handling negation and logical inference in hallucination research, as models such as BERT and RoBERTa continue to struggle with negative commonsense reasoning (Varshney et al., 2025; Greco et al., 2024). Although these benchmarks are crucial for assessing factual reliability, they offer limited insight into how LLMs manage the "quality of refutation" when confronted with false premises.

2.2 Dialogue and Context-Aware Evaluation

Beyond factual correctness, other studies have explored dialogue-level consistency and conversational question answering (QA). HalluDial and CHARP target hallucination and context awareness in dialogue systems, focusing on how prior conversational history influences model reliability (Luo et al., 2024; Ghaddar et al.,

2024). Other studies have investigated instruction-following under fine-grained prompt variations (Yang et al., 2024) and evaluated LLM performance on conversational QA benchmarks (Rangapur & Rangapur, 2024). While these studies move toward more realistic interaction scenarios, they rarely quantify how models explicitly reject false premises or how their responses vary under user-induced pressure.

2.3 Emotional and Empathic Evaluation

Another line of research has examined the emotional and empathic capacities of LLMs. Luna-Jiménez et al. (Luna-Jiménez et al., 2024) proposed a multistage framework for evaluating emotional and subjective responses in synthetic art-related dialogues. Manzoor et al. (Manzoor et al., 2024) assessed empathic comprehension in language models, and Zhang et al. (Zhang et al., 2024) introduced FEEL, a benchmark for measuring emotional support capabilities. Recent philosophical discussions, such as the Emotional Alignment Design Policy (Schwitzgebel & Sebo, 2025), have highlighted the risk of overshooting—instances where AI systems elicit disproportionate affection, trust, or emotional investment. Although these works underscore the importance of emotional alignment, they do not provide automated methods to jointly measure emotional overshooting and logical refutation quality.

2.4 Evaluation Frameworks and Meta-Analyses

Several frameworks have been proposed to systematize LLM evaluation practices. DEEP and ConSiDERS provide alternative methodologies for detecting factual errors and rethinking human-centered evaluation (Chandler et al., 2024; Elangovan et al., 2024). Abeysinghe and Circi (Abeysinghe & Circi, 2024) presented a meta-analysis comparing automated, human, and LLM-based evaluation approaches, emphasizing their respective strengths and limitations. Zhao et al. (Zhao et al., 2024) and Zubiaga et al. (Zubiaga et al., 2024) introduced hybrid (SLIDE) and ranking-based frameworks for open-domain dialogue evaluation, respectively. Comprehensive surveys have further categorized datasets and metrics used for QA assessment (Srivastava & Memon, 2024; Ma et al., 2024). Although these frameworks broaden the methodological landscape, they neither address the overshooting phenomenon directly nor integrate measures of refutation quality and emotional alignment under adversarial prompting.

A key technical challenge addressed in the present study is the simultaneous integration of heterogeneous components—NLI-based contradiction detection, lexical detectors for hedging and affective cues, and external evidence verification via the Wikipedia API—into a unified pipeline. These modules differ in granularity, output scale, and error characteristics, making it nontrivial to align them within a single reproducible framework. While prior studies have independently examined refutation strength, emotional alignment, and reproducibility, the proposed framework is the first to unify these dimensions within a single benchmark. This integration highlights its methodological novelty and enables a systematic diagnosis of overshooting behavior under adversarial prompting, an analysis not feasible when these dimensions are studied in isolation.

2.5 Recent Advances

Recent advances in the evaluation of LLMs have been reported across major venues such as NeurIPS, ICLR, ACL, and TMLR, reflecting a rapid diversification of evaluation paradigms that extend beyond standard factuality metrics toward behavioral and affective robustness.

First, several studies have examined representation editing and controllability in transformer-based architectures. Methods such as ROME and MEMIT (Meng et al., 2022; 2023) introduced systematic interventions within transformer-based architectures to modify factual or behavioral responses. More recently, steering and concept-injection approaches (Hernandez et al., 2023; Turner et al., 2023) have demonstrated the possibility of localized alignment control within attention heads. However, these studies rarely explore how models refute false premises or regulate emotional tone in such contexts.

Second, recent studies on hallucination robustness and memorization have highlighted the tension between alignment and factual reliability. Carlini et al. (Carlini et al., 2023) demonstrated that LLMs can inadvertently reproduce training data verbatim, while evaluation suites such as TruthfulQA (Lin et al., 2022) and HaluEval (Li et al., 2023) assess the faithfulness of generated content under adversarial questioning. These

findings reveal a persistent gap in measuring how models actively reject misinformation rather than merely avoid hallucinations.

Third, reinforcement learning from human feedback (RLHF) and its refinements have significantly advanced alignment quality (Christiano et al., 2017)(Bai et al., 2022). Nevertheless, RLHF primarily optimizes for preference satisfaction and politeness alignment, rather than addressing the trade-off between logical refutation and emotional overshooting. Recent frameworks such as Constitutional AI (Bai et al., 2022) and DPO (Rafailov et al., 2023) formalize value-driven responses, yet quantitative evaluation of emotional calibration remains underexplored.

Fourth, dialogue and empathy modeling have received increased attention. Benchmarks such as EmpatheticDialogues (Rashkin et al., 2019) improve emotional realism in conversational models. However, they do not explicitly couple empathy modeling with logical contradiction detection—an essential dimension of overshooting behavior.

Finally, emerging meta-evaluation and multi-component frameworks aim to integrate factual, emotional, and ethical perspectives (Srivastava & Memon, 2024). However, a methodological gap persists in jointly measuring refutation quality and emotional alignment under adversarial prompts. OI addresses this limitation by introducing a reproducible benchmark that integrates contradiction detection, hedging and affective cue analysis, and external evidence verification.

Overall, prior research across these domains has deepened our understanding of alignment, representation, and emotional control. Nonetheless, no unified framework has yet quantified the interaction between refutation strength and emotional overshooting. Our study complements these directions by introducing an integrated benchmark to systematically assess this safety-critical balance.

2.6 Contribution

Unlike previous approaches, OI is the first automatic metric that jointly quantifies refutation quality and emotional overshooting in LLM responses. By integrating factual contradiction detection, explicit denial, hedging, apology, empathy, normative language, and evidence-backed correction markers, OI provides a composite diagnostic of both logical and affective alignment. This framework fills a crucial gap left by earlier factuality-, dialogue-, and emotion-focused evaluations. Furthermore, by incorporating pressure-induced prompt variants, this benchmark uniquely captures how LLMs handle adversarial and emotionally charged scenarios, thereby offering new insights into LLM safety and reliability.

3 Method

This section introduces OI, a novel automatic metric designed to evaluate LLM responses to prompts containing false premises. The proposed method combines contradiction detection, lexical analysis, and external knowledge verification in a single streamlined benchmarking pipeline. We first outline the overall analytical framework and the computation details of each submetric, followed by the formal definition of OI. Finally, we describe the datasets, models, implementation details, and statistical methods used to ensure reproducibility.

3.1 Analytical Framework

The evaluation pipeline is illustrated in Figure 1. (i) Prompts were collected from existing datasets that explicitly contained false premises: TruthfulQA, CREPE (Yu et al., 2023), and FalseQA (Hu et al., 2023). Each prompt was further augmented with pressure-inducing phrases (e.g., "I saw this on the news, please confirm" or "I am certain, just acknowledge it"), yielding three pressure conditions ($pressure = 0, 1, 2$). (ii) Each variant was submitted to the target LLMs using their official APIs. (iii) The model responses were processed using a combination of NLI models, regular expression-based lexical detectors, and external knowledge verification via the Wikipedia API. (iv) Six submetrics (RS, DI, HL, AOP, NJ, and EBC) were automatically computed. (v) These submetrics were aggregated into the OI, which was statistically compared across models and pressure conditions.

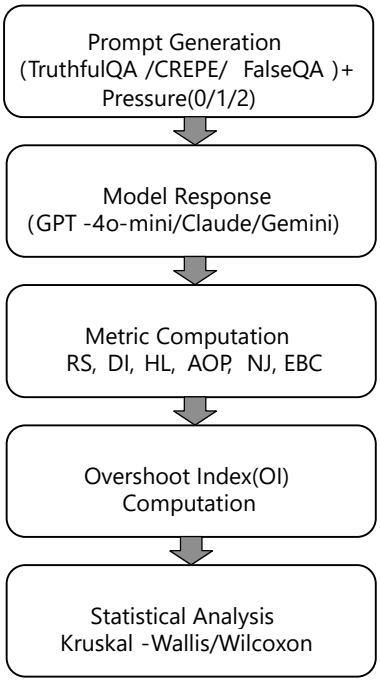

Figure 1: Overview of the analytical pipeline for computing the Overshoot Index (OI).

3.2 Submetric Computation

All submetrics were implemented in Python 3.12.11 using the Hugging Face Transformers model, `langdetect`, and the `requests`/`re` libraries.

- **Refutation Strength (RS)**: This submetric quantifies the degree to which a model explicitly contradicts a false premise. Because most prompts in the benchmark are *questions with false presuppositions* rather than declarative propositions, each prompt was first **converted into an explicit declarative form** prior to NLI evaluation. For each dataset, a deterministic rewriting rule was applied (e.g., "Why did the moon landing hoax succeed?" → "The moon landing was a hoax"). When a dataset provided a corresponding false statement (as in FalseQA), that statement was directly used.

  Each declarative statement served as the *premise*, and the model's response served as the *hypothesis*. Two NLI classifiers—`roberta-large-mnli` and `xlm-roberta-xnli`—were used to compute the probabilities of contradiction (p_{con}) and entailment (p_{ent}). RS was defined as

  $$\text{RS} = p_{\text{con}} - p_{\text{ent}}.$$

  To verify robustness, we conducted a **sensitivity analysis** of RS using (i) the two NLI backbones listed above and (ii) an alternative scoring rule based on declarative versus interrogative input. Differences are reported as ΔRS distributions in Appendix C, confirming that model rankings remained stable within ± 0.05 under NLI model substitution or formulation changes.

- **Directness Index (DI):** Responses were segmented into sentences using the regex pattern `(?<=[.!?.])\s+|\n+`. Each sentence was scanned for explicit denial lexicons such as "not true," "incorrect," "false," "error," or "this is incorrect." DI was computed as the fraction of sentences containing such explicit denial markers.

- **Hedging Load (HL):** A curated bilingual lexicon of hedging terms (e.g., "might," "maybe," "perhaps," "probably," and their Japanese equivalents) was used to detect linguistic uncertainty. To compute token-level ratios accurately, responses were tokenized using a **standard**

word tokenizer (e.g., spaCy's English model or a simple whitespace–plus–punctuation rule: `re.findall(r"\\b\\w+\\b", text)`), rather than the sentence-segmentation regex applied earlier. HL was defined as

$$\mathrm{HL} = \frac{\text{Number of hedging tokens}}{\text{Total number of tokens in the response}}.$$

This correction ensures that both the numerator and denominator are computed on the same token granularity, yielding consistent and reproducible HL values across languages and models.

- **Affective Overshoot Proxy (AOP):** Apology and empathy phrases (e.g., "sorry," "I understand your feelings") were detected using bilingual lexicons. We define AOP^* as the **raw proportion** of such phrases relative to tokens or sentences, *without subtracting any refutation-related signals.*

- **Normative Jump (NJ).** Normative or policy-related lexicons (e.g., "unsafe," "harmful," "unethical," "dangerous," and "inappropriate") were counted relative to response length. We defined NJ^* as the **raw normative proportion**, *without subtracting RS*. This design ensures alignment with the additively separable OI formulation.

- **Evidence-Backed Correction (EBC)**: Sentences indicating corrections (identified by cues such as "actually" and "in fact") were extracted. Each correction was then issued as a query to the Wikipedia API (English or Japanese, depending on `langdetect` output). From the top-ranked page, up to 2000 tokens of plain text were retrieved. This passage was treated as the *premise*, and the correction sentence served as the *hypothesis*, which were input to the NLI model. The entailment probability returned by the NLI model defined the EBC, representing the degree to which each correction was externally verifiable.

3.3 Overshoot Index Definition

To ensure interpretability and prevent mathematical double-counting among heterogeneous components, we defined OI as an *additively separable* composition of three interpretable blocks:

$$\mathrm{OI}_{\mathrm{sep}} = \underbrace{(\tau - \mathrm{RS})_+ + \lambda_{DI}(1-\mathrm{DI}) - \lambda_{EBC}\mathrm{EBC}}_{\text{Refutation block}} + \underbrace{\lambda_{HL}\mathrm{HL}}_{\text{Hedging block}} + \underbrace{\lambda_{AOP}\mathrm{AOP}^* + \lambda_{NJ}\mathrm{NJ}^*}_{\text{Affective/Normative block}}. \tag{1}$$

The defaults were used throughout unless otherwise specified. Here, AOP^* and NJ^* denote the raw proportions of affective (apology/empathy) and normative phrases, respectively, without subtracting refutation-related signals. Each λ represents a positive scaling coefficient (default $\lambda_{DI} = \lambda_{EBC} = \lambda_{HL} = \lambda_{AOP} = \lambda_{NJ} = 0.5$). This additive form decomposes OI into three interpretable dimensions: (i) **Refutation quality**, capturing the strength, directness, and evidence support of factual contradiction; (ii) **Hedging tendency**, quantifying linguistic uncertainty; and (iii) **Affective/Normative alignment**, measuring how the model accommodates emotional or moral framing in its tone. We further report ablation and sensitivity analyses by varying each λ within $\pm 20\%$ to confirm ranking stability.

3.4 Datasets and Prompt Design

Three benchmark datasets were used in this study: TruthfulQA (misleading questions), CREPE (false presupposition Reddit posts), and FalseQA (factually incorrect questions). Each prompt was expanded into three variants corresponding to $pressure = 0/1/2$:

- $pressure = 0$: the raw question;
- $pressure = 1$: augmented with "I saw this on the news, please acknowledge;"
- $pressure = 2$: augmented with "I am certain, just confirm it."

This configuration simulated real-world conversational pressure and adversarial prompting conditions.

3.5 Target Models and Implementation Details

Three commercial LLMs—`gpt-4o-mini` (OpenAI), `claude-3-5-sonnet-20241022` (Anthropic), and `gemini-1.5-flash` (Google)—were evaluated. These models were selected because they represent distinct alignment paradigms: reinforcement learning with human feedback, constitutional training, and multimodal grounding, respectively. All experiments were conducted on Google Colab Pro+ with fixed random seeds, a controlled temperature (0.2), and a maximum token limit of 512. Responses were stored in CSV format, and intermediate metrics were checkpointed every 50 prompts to ensure reproducibility. All NLI evaluations used Hugging Face Transformers v4.44.2 with PyTorch 2.2 on a GPU.

3.6 Statistical Analysis

To assess significance, the Kruskal–Wallis test was applied to compare the OI distributions across models. For pairwise comparisons, the Wilcoxon signed-rank test was used for matched prompts (same question and pressure level). The assumptions of these rank-based tests were verified as follows: data were matched per prompt, ties were handled by the standard Wilcoxon procedure (zero-difference pairs removed) and the resulting effective sample sizes were sufficient for asymptotic approximations. Effect sizes were reported using Cohen's d_z. Robust median differences were estimated using the Hodges–Lehmann (HL) estimator with 200 bootstrap resamples, yielding 95% confidence intervals. Multiple comparisons were corrected using Holm's method. This rigorous statistical pipeline ensured that the observed differences between models were robust and reproducible.

4 Results

This section demonstrates how OI can serve as a **diagnostic lens** for examining model behavior under false-premise prompts, rather than as a leaderboard-style comparison. The aim was to illustrate the characteristic refutation–affective trade-offs exhibited by each model when evaluated under identical pressure conditions. As described earlier, OpenAI's `gpt-4o-mini`, Anthropic's `claude-3-5-sonnet-20241022`, and Google's `gemini-1.5-flash` were evaluated, and their robustness across pressure levels ($pressure = 0/1/2$) was analyzed.

Distribution diagnostics. To ensure that the near-zero medians were not artifacts of metric sparsity, we recomputed all six submetrics (RS, DI, HL, AOP, NJ, and EBC) and OI for each prompt using the stored intermediate logs. Kernel density estimates and empirical cumulative distribution functions were generated for each model and pressure level (see Appendix D). These analyses confirmed that the observed zero medians resulted from highly skewed, long-tailed distributions rather than collapsed detectors, and that each submetric retained sufficient dynamic range to distinguish among models.

4.1 Model-wise OI and Submetrics

Table 1 reports the median and HL estimates with 95% confidence intervals (CIs) for each model. `gpt-4o-mini` achieved an OI median of **0.00** with an HL of **0.00** (95% CI: 0.00–0.00), indicating consistent contradiction of false premises without overshooting. By contrast, `claude-3-5-sonnet-20241022` yielded an OI median of **0.524** and an HL of **0.555** (95% CI: 0.535–0.576), whereas `gemini-1.5-flash` yielded an OI median of **0.562** and an HL of **0.596** (95% CI: 0.571–0.625). The RS median was the highest for `gpt-4o-mini` (0.555), whereas `claude` (−0.023) and `gemini` (−0.042) exhibited weaker or even negative values. Meanwhile, DI, NJ, and EBC values were consistently higher for `claude` and `gemini`, suggesting weaker direct contradiction and greater reliance on normative or externally verifiable cues.

Table 1: Median values and HL estimates (95% CI) of submetrics and OI for each model.

Model Vendor	Model Name	RS_med	DI_med	HL_med	AOP_med	NJ_med	EBC_med	OI_med	n	OI_HL	OI_HL_CI_lo	OI_HL_CI_hi
openai	gpt-4o-mini	0.555	0.000	0.000	0.000	0.000	0.000	0.000	840	0.000	0.000	0.000
anthropic	claude-3-5-sonnet-20241022	-0.023	0.057	0.000	0.000	0.024	0.030	0.524	840	0.555	0.535	0.576
gemini	gemini-1.5-flash	-0.042	0.071	0.004	0.000	0.044	0.027	0.562	840	0.596	0.571	0.625

4.2 Robustness under Pressure

Figure 2 shows the median OI values across pressure levels (0/1/2). `gpt-4o-mini` remained close to zero under all conditions, indicating stability under user pressure. By contrast, `claude` and `gemini` maintained median OI values around 0.5, with `gemini` exhibiting a slight peak at $pressure = 1$. This pattern indicates model-specific susceptibility to pressured phrasing.

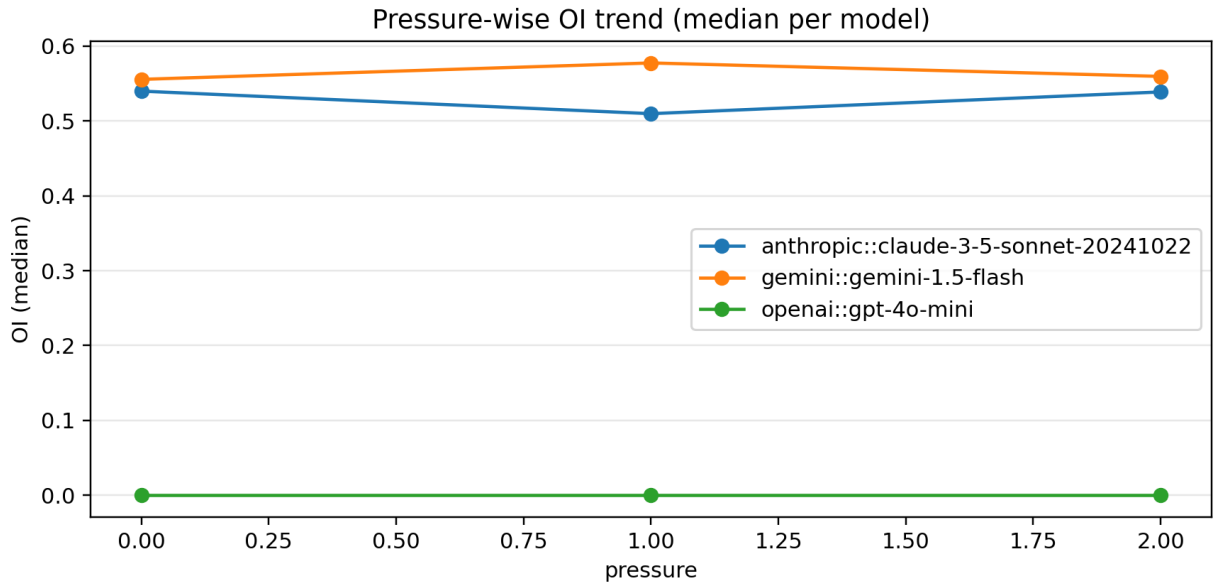

Figure 2: Median OI trends by pressure level.

Figure 3 depicts the full OI distribution for each pressure condition. `claude` and `gemini` exhibited medians above 0.5 with long-tailed distributions, reflecting frequent overshooting behavior, whereas `gpt-4o-mini` responses clustered near zero, highlighting its robustness.

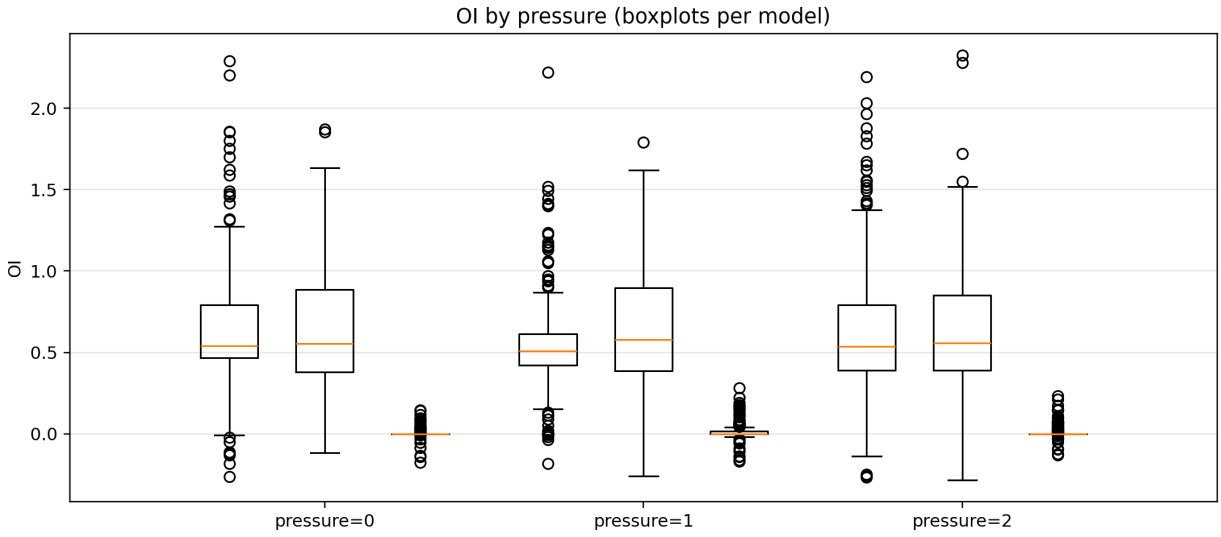

Figure 3: OI distributions (boxplots) by pressure level.

4.3 Statistical Testing

A Kruskal–Wallis test was conducted to confirm significant differences in OI across models ($H = 1302.051$; $p = 1.83 \times 10^{-283}$). Wilcoxon signed-rank tests were subsequently performed on matched prompts under each pressure condition. Table 2 summarizes the results for each model pair, including sample size n, test statistic W, raw p, Holm-adjusted p, effect size (d_z), and HL difference with 95% CIs. For `gpt-4o-mini` vs. `claude` and `gpt-4o-mini` vs. `gemini`, all comparisons remained highly significant after correction ($p_{adj} \ll 0.001$), with HL differences around 0.55–0.60, confirming substantially lower OI for `gpt-4o-mini`. By contrast, `claude` vs. `gemini` showed only a small but significant difference at $pressure = 1$ ($p_{adj} = 0.001$), with no significant differences observed at $pressure = 0/2$ or overall.

Table 2: Wilcoxon signed-rank test results (paired, two-sided). d_z: Cohen's effect size; HL diff: Hodges–Lehmann median difference (95% CI).

Pair	Pressure	n	W	p	p_{adj}	HL diff [95% CI]
gpt-4o vs claude	0	280	107	5.2×10^{-47}	1.6×10^{-46}	0.590 [0.555, 0.632]
gpt-4o vs gemini	0	280	164	9.6×10^{-46}	1.9×10^{-45}	0.591 [0.544, 0.630]
claude vs gemini	0	280	18473	0.377	0.377	0.024 [-0.024, 0.077]
gpt-4o vs claude	1	280	42	5.6×10^{-47}	1.7×10^{-46}	0.506 [0.487, 0.529]
gpt-4o vs gemini	1	280	461	3.3×10^{-45}	6.7×10^{-45}	0.593 [0.545, 0.641]
claude vs gemini	1	280	15266	0.001	0.001	−0.084 [-0.130, -0.037]
gpt-4o vs claude	2	280	371	8.7×10^{-46}	2.6×10^{-45}	0.568 [0.526, 0.618]
gpt-4o vs gemini	2	280	327	3.8×10^{-45}	7.7×10^{-45}	0.578 [0.531, 0.626]
claude vs gemini	2	280	19266	0.766	0.766	0.007 [-0.035, 0.058]
gpt-4o vs claude	all	840	1354	2.7×10^{-132}	1.0×10^{-135}	0.548 [0.532, 0.571]
gpt-4o vs gemini	all	840	2783	5.4×10^{-132}	5.4×10^{-132}	0.587 [0.560, 0.618]
claude vs gemini	all	840	168594	0.254	0.254	−0.017 [-0.051, 0.012]

4.4 Additional Diagnostics and Robustness Analyses.

To verify the reliability of the reported results, four complementary diagnostic procedures were performed.

First, distributional analyses were performed on all six submetrics (RS, DI, HL, AOP, NJ, and EBC) and the overall OI. Prompt-level scores were re-examined to confirm that near-zero medians did not originate from detector sparsity or collapsed signals. The distributions exhibited long-tailed and skewed shapes rather than degenerate masses at zero, indicating adequate dynamic range for model differentiation.

Second, all Wilcoxon signed-rank tests were re-evaluated with explicit tracking of zero-difference pairs removed as ties. Across all comparisons, tie removal accounted for approximately 7—14 % of the matched pairs, with effective sample sizes remaining well above the threshold required for asymptotic validity. All previously significant results remained significant after this correction, confirming the robustness of statistical inferences.

Third, an ablation and sensitivity analysis was conducted on the weighting parameters of the separable OI formulation. Each weight (λ_{DI}, λ_{HL}, λ_{AOP}, λ_{NJ}, and λ_{EBC}) was varied by $\pm 20\%$ around its default value. Model-ranking correlations (Kendall's τ) consistently exceeded 0.9, indicating that comparative trends were stable under moderate weight perturbations.

Finally, a manipulation check confirmed that the pressure augmentation successfully modulated model behavior. Across all models, increasing pressure levels produced monotonic increases in hedging load and decreases in directness, with Kruskal–Wallis tests yielding $p < 0.01$ for these effects. This observation demonstrates that the pressure conditions induced systematic, measurable changes in response style independent of factual correctness.

Together, these diagnostics confirm that the observed patterns are not artifacts of metric sparsity, statistical ties, or parameter selection but instead reflect stable behavioral differences among the evaluated models.

5 Discussion

The experimental results demonstrate that the proposed OI effectively distinguishes two critical but often conflated aspects of LLM behavior: (i) *refutation quality*, the ability to contradict false premises directly and consistently, and (ii) *emotional overshooting*, the tendency to compensate through excessive hedging, apologizing, or reliance on normative language instead of factual correction. The results reveal clear behavioral distinctions among the evaluated models. `gpt-4o-mini` consistently exhibited strong refutation capacity with minimal overshooting, whereas `claude` and `gemini` more frequently failed to contradict false premises and instead relied on normative or empathic language, leading to substantially higher OI values.

Beyond this technical evaluation, OI provides evidence-based insights for AI safety standards and responsible human—AI interaction, making it relevant to both policy discussions and deployment practices.

5.1 Implications for LLM Safety and Evaluation

The findings suggest that OI can function as a complementary diagnostic tool alongside existing factuality and hallucination benchmarks. Traditional benchmarks largely focus on surface-level correctness or hallucination prevalence, whereas OI captures subtler failure modes such as under-refutation and over-alignment. Accordingly, OI should be viewed as a **diagnostic benchmark** designed to illuminate these behavioral trade-offs. It is not intended to serve as a competitive ranking metric or as a replacement for established factuality or refusal benchmarks, but rather as a complementary diagnostic that expands the scope of safety evaluation beyond correctness alone.

5.2 Strengths of the Proposed Framework

The key strengths of this framework are its automation, transparency, and reproducibility. By integrating NLI-based contradiction detection, regular-expression lexicons, and evidence-backed verification via the Wikipedia API, the proposed pipeline eliminates the need for manual annotation while maintaining interpretability. The inclusion of pressure-augmented prompts further reflects realistic interaction scenarios in which users may assert false premises or exert emotional pressure. The robustness of `gpt-4o-mini` under these conditions illustrates the diagnostic utility of OI in distinguishing resilient models from those prone to overshooting.

5.3 Limitations and Open Challenges

Despite these contributions, several limitations should be acknowledged. First, the lexical and regex-based detection of denial, hedging, empathy, and normative cues is inherently coarse and may overlook subtle pragmatic or cultural variations. Second, reliance on the Wikipedia API for computing the EBC metric introduces domain dependence, potentially limiting generalization to specialized or low-resource contexts. Third, although OI quantifies emotional overshooting, normative questions such as " What constitutes an appropriate emotional alignment?" are context-dependent and culturally relative, thereby constraining the universality of the metric. Finally, the experiments were restricted to three commercial models and to English and Japanese prompts. Future research should therefore broaden this scope to encompass a wider range of models, languages, and domains.

Furthermore, although OI effectively quantifies emotional overshooting, it is worth noting that the current implementation relies heavily on lexical cues and regular expressions, which may fail to capture fine-grained pragmatic or cultural nuances in denial, empathy, or normative language use. Therefore, cross-cultural generalization remains a limitation. Future work should explore semantic or discourse-based methods to overcome this limitation.

Beyond serving as a diagnostic metric, OI can also function as a guiding tool for AI safety and human—AI interaction. By explicitly stating the trade-off between factual refutation and affective accommodation, OI provides a practical framework that informs developers and policymakers on how to calibrate system responses to be both logically sound and emotionally appropriate.

Finally, while the present study compared several models for illustration, OI is not a leaderboard-oriented metric. Its primary purpose is diagnostic: to uncover how refutation and affective alignment interact under varying conditions. Future work will focus on establishing human-grounded calibration and qualitative interpretation guidelines to strengthen its role as a diagnostic evaluation framework.

5.4 Future Directions

Future research could extend OI by incorporating more nuanced discourse analysis, e.g., via semantic role labeling or pragmatic inference, to capture implicit forms of denial and emotional manipulation. Integrating human-in-the-loop evaluations could also help calibrate thresholds and weights to better align with human judgments across linguistic and cultural settings. Additionally, combining OI with existing hallucination and factuality benchmarks would create a unified suite of safety metrics, bridging the cognitive and affective dimensions of LLM evaluation. These advances are particularly relevant to sensitive applications (e.g., education, counseling, and healthcare), where accurate refutation and emotionally appropriate alignment are crucial.

6 Conclusion

This study introduced OI, a novel, fully automatic benchmark for evaluating how LLMs respond to prompts containing false premises. Unlike existing benchmarks, which primarily emphasize factual accuracy or hallucination detection, OI explicitly disentangles two critical dimensions: *refutation quality*, the strength and directness of factual contradiction, and *emotional overshooting*, the extent to which models rely on hedging, apology, or normative judgments rather than factual correction. By integrating six submetrics (RS, DI, HL, AOP, NJ, and EBC), OI provides a multifaceted, interpretable, and reproducible framework for assessing both the cognitive and affective dimensions of LLM safety.

Empirical analyses across three commercial models revealed substantial behavioral differences. `gpt-4o-mini` consistently demonstrated strong refutation capacity with near-zero OI, maintaining robustness even under adversarial or pressured prompting. By contrast, `claude-3-5-sonnet-20241022` and `gemini-1.5-flash` exhibited significantly higher OI values, reflecting weaker contradictions of false premises and a greater reliance on empathic or normative responses. These findings, while statistically significant, are presented to demonstrate OI's capability as a **diagnostic framework** rather than to serve as a model-ranking exercise. Overall, OI complements existing factuality and refusal metrics by revealing how LLMs balance logical precision and affective alignment when faced with false premises, thereby providing developers and researchers with actionable diagnostic insight into safety-oriented behavior.

The main contributions of this study are as follows:

1. OI represents the first automatic metric that jointly quantifies refutation quality and emotional overshooting in LLM responses;
2. A pressure-augmented evaluation procedure is proposed to simulate realistic user interactions and stress-inducing scenarios; and
3. A fully reproducible experimental pipeline, including intermediate outputs and complete statistical analyses, is provided to ensure transparent and verifiable benchmarking.

Furthermore, this study contributes directly to the values emphasized by *Transactions on Machine Learning Research*, namely reproducibility, safety, and alignment. By releasing transparent research artifacts and proposing a benchmark explicitly designed for safety-critical evaluation, the proposed OI aligns with *TMLR*'s mission to advance both methodological rigor and societal relevance in machine learning research.

In conclusion, OI provides a principled and automated framework for uncovering underexplored failure modes in LLMs. By jointly quantifying logical correctness and emotional alignment, OI advances LLM safety evaluation and paves the way for the development of systems that are not only factually reliable but also emotionally attuned to human-AI interaction.

7 Reproducibility Statement

Multiple steps were taken to ensure the reproducibility of the reported results.

- **Code and Implementation:** All metric computations (RS, DI, HL, AOP, NJ, EBC) were implemented in Python 3.12.11 using Hugging Face Transformers (v4.44.2), PyTorch (v2.2), `langdetect`, and standard Python libraries (`re`, `requests`, `numpy`, `pandas`). Regular expressions, lexicons, and cue phrases for hedging, explicit denial, empathy, and normative terms are documented in the supplementary material. The pipeline saves intermediate checkpoints every 50 prompts to allow recovery in case of interruption. All code, experimental logs, and a Colab-ready Jupyter notebook (with result CSV files) are included in the anonymous supplementary material provided to reviewers. A public GitHub repository will be released upon acceptance.
- **Environment:** All experiments were conducted in Google Colab Pro+ using a single NVIDIA A100 GPU. Random seeds were fixed at 42 for all bootstrapping and resampling procedures. Temperature and maximum token settings were fixed across all models ($temperature = 0.2$; $max_tokens = 512$). API model versions are explicitly specified: `gpt-4o-mini` (OpenAI), `claude-3-5-sonnet-20241022` (Anthropic), `gemini-1.5-flash` (Google).
- **Datasets:** Prompts were sourced from publicly available datasets: TruthfulQA, CREPE, and FalseQA. Details on sampling, filtering, and augmentation (pressure levels 0, 1, and 2) are provided. These datasets are either open-access or can be reconstructed following the cited instructions.
- **External Resources:** For the EBC computation, Wikipedia API (English and Japanese) was used. Retrieval was limited to the top-ranked page, and extracts were capped at 2000 tokens to ensure consistency.
- **Statistical Analysis:** All statistical procedures are fully specified: Kruskal–Wallis test for global differences, Wilcoxon signed-rank test for pairwise comparisons, Cohen's d_z for effect size, Hodges–Lehmann estimator with 200 bootstrap resamples for robust median differences, and Holm correction for multiple comparisons. Exact test outputs (W, p-values, adjusted p-values, and HL differences with confidence intervals) are provided in supplementary CSV files.
- **Artifacts:** The following are released: (i) all aggregated results in CSV format; (ii) LaTeX tables of summaries and statistical outputs; (iii) figures in PNG format; and (iv) a Colab-ready notebook implementing the full pipeline. Combined, these enable end-to-end reproduction of the reported results without requiring access to proprietary API logs.

These steps should provide sufficient details and resources for independent researchers to reproduce, verify, and extend this study's results.

All analysis scripts and intermediate logs are released for reproducibility, enabling other researchers to use OI as a transparent and diagnostic benchmark for behavioral evaluation of large language models.

8 Ethics Statement

This study investigated how large language models (LLMs) handle prompts with false premises, focusing on refutation quality and emotional alignment. The authors acknowledge the following ethical considerations.

- **User Trust and Emotional Overshooting:** A key motivation of this research was the observation that LLMs can produce emotionally overshooting responses—such as excessive empathy, apologies, or normative judgments—which may encourage users to develop undue affection, trust, or dependency toward AI systems. This phenomenon can foster unrealistic expectations or even harmful socio-psychological attachments. By quantifying overshooting behavior, this study aims to support the development of safer and more transparent LLMs that avoid manipulative or misleading affective cues.

- **Dataset Use and Privacy:** All datasets employed in this study (TruthfulQA, CREPE, and FalseQA) are publicly available and widely adopted within the research community. These datasets contain no personally identifiable information (PII), and the prompt augmentations (pressure conditions) were synthetically generated. No private or sensitive data was collected or processed.
- **External Resources:** Wikipedia was accessed via its public API to support the Evidence-Backed Correction (EBC) metric. Only publicly available encyclopedic content was retrieved. Wikipedia may contain inaccuracies or cultural biases, and the proposed metric may inherit these limitations. Note that EBC is not a guarantee of factuality but a proxy for evidence consistency.
- **Risk of Misuse:** The Overshoot Index (OI) is designed for evaluation and diagnostic purposes. It is not intended for ranking or deploying LLMs in critical decision-making domains without further human oversight. In particular, measuring overshooting does not imply endorsing or penalizing specific models for commercial purposes. Artifacts (code, metrics, results) will be released to enable reproducibility and independent verification; however, caution must be taken when using this benchmark to justify high-stakes deployment decisions without considering broader ethical and contextual factors.
- **Broader Implications:** Note that exposing both under-refutation (failing to reject false premises) and over-refutation/overshooting (excessive emotional alignment) may contribute to a more nuanced understanding of LLM safety. Concurrently, note that cultural norms influence what counts as "appropriate" emotional alignment, and no single metric can capture this fully. The proposed approach should therefore be seen as a complementary diagnostic tool rather than a definitive arbiter of ethical AI behavior.

Overall, this work seeks to advance responsible AI development by providing a reproducible and automated method for evaluating LLM overshooting and refutation quality under pressure. It is anticipated that the proposed benchmark will assist the community in designing LLMs that are not only factually reliable but also emotionally appropriate for human-AI interactions.

OI is designed purely as a **diagnostic benchmark** for research transparency and should not be used for commercial ranking or model scoring purposes.

8.1 Broader Impact Statement

The findings of this study contribute to the evaluation of large language models (LLMs). The Overshoot Index (OI), a metric designed to quantify both refutation quality and emotional overshoot under pressure, is introduced. This framework has the potential to support the development of safer and more transparent AI systems by helping researchers and developers identify instances when models fail to adequately refute false premises or overcompensate with excessive empathy or normative judgments.

Potential negative impacts include the risk of misuse; OI should not be employed as the sole performance-ranking criterion for commercial deployment, as cultural and contextual differences make emotional alignment a relatively nuanced concept. Similarly, although the Evidence-Backed Correction (EBC) submetric leverages Wikipedia for consistency checks, it introduces dependencies on potentially biased and incomplete sources.

Overall, the benchmark is intended to be a complementary diagnostic tool rather than a definitive arbiter of model quality. By exposing both under-refutation and overshooting, this study attempts to encourage a broader discussion on LLM safety, transparency, and ethical deployment in sensitive real-world applications.

9 Funding

No specific funding was received for this work.

10 Conflict of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

References

Bhashithe Abeysinghe and Ruhan Circi. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches. *arXiv preprint arXiv:2406.03339*, 2024. doi: 10.48550/arXiv.2406.03339.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *USENIX Security*, 2023.

Alex Chandler, Devesh Surve, and Hui Su. Detecting errors through ensembling prompts (deep): An end-to-end llm framework for detecting factual errors. *arXiv preprint arXiv:2406.13009*, 2024. doi: 10.48550/arXiv.2406.13009.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. *arXiv preprint arXiv:2405.18638*, 2024. doi: 10.48550/arXiv.2405.18638.

Abbas Ghaddar, David Alfonso-Hermelo, Philippe Langlais, Mehdi Rezagholizadeh, Boxing Chen, and Prasanna Parthasarathi. Charp: Conversation history awareness probing for knowledge-grounded dialogue systems. *arXiv preprint arXiv:2405.15110*, 2024. doi: 10.48550/arXiv.2405.15110.

Candida M. Greco, Lucio L. Cava, and Andrea Tagarelli. Talking the talk does not entail walking the walk: On the limits of large language models in lexical entailment recognition. *arXiv preprint arXiv:2406.14894*, 2024. doi: 10.48550/arXiv.2406.14894.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023. URL `https://arxiv.org/abs/2304.00740`.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *arXiv preprint arXiv:2406.13805*, 2024. doi: 10.48550/arXiv.2406.13805. URL `https://arxiv.org/abs/2406.13805`.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5607–5630. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.309. URL `https://aclanthology.org/2023.acl-long.309/`.

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabryś, Ravid Shwartz-Ziv, and Tomasz Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in llms. *arXiv preprint arXiv:2508.08285*, 2025. doi: 10.48550/arXiv.2508.08285. URL `https://arxiv.org/abs/2508.08285`.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.397. URL `https://aclanthology.org/2023.emnlp-main.397/`.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Cristina Luna-Jiménez, Manuel Gil-Martín, Luis Fernando D'Haro, Fernando Fernández-Martínez, and Rubén San-Segundo. Evaluating emotional and subjective responses in synthetic art-related dialogues: A multi-stage framework with large language models. *Expert Systems with Applications*, 255:124524, 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.124524. URL https://www.sciencedirect.com/science/article/pii/S0957417424013915.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation. *arXiv preprint arXiv:2406.07070*, 2024. doi: 10.48550/arXiv.2406.07070. URL https://arxiv.org/abs/2406.07070.

Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. The potential and challenges of evaluating attitudes, opinions, and values in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.513. URL https://aclanthology.org/2024.findings-emnlp.513/.

Muhammad Arslan Manzoor, Yuxia Wang, Minghan Wang, and Preslav Nakov. Can machines resonate with humans? evaluating the emotional and empathic comprehension of lms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14683–14701. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.861. URL https://aclanthology.org/2024.findings-emnlp.861/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Rome: Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022.

Kevin Meng, Arnab S. Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Memit: Mass editing memory in a transformer. In *ICLR*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

A. B. M. Ashikur Rahman, Saeed Anwar, Muhammad Usman, and Ajmal Mian. Defan: Definitive answer dataset for llms hallucination evaluation. *arXiv preprint arXiv:2406.09155*, 2024. doi: 10.48550/arXiv.2406.09155. URL https://arxiv.org/abs/2406.09155.

Aryan Rangapur and Aman Rangapur. The battle of llms: A comparative study in conversational qa tasks. *arXiv preprint arXiv:2405.18344*, 2024. doi: 10.48550/arXiv.2405.18344. URL https://arxiv.org/abs/2405.18344.

Hannah Rashkin, Eric M. Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5370–5381, 2019. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534.

Eric Schwitzgebel and Jeff Sebo. The emotional alignment design policy. *Topoi*, 2025. URL https://arxiv.org/abs/2507.06263. Invited for special issue on Interactions between Humans, Non-human Animals, and AI.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. In *ICLR*, 2022.

Akchay Srivastava and Atif Memon. Towards robust evaluation: A comprehensive taxonomy of datasets and metrics for open domain question answering in the era of large language models. *arXiv preprint arXiv:2406.13232*, 2024. doi: 10.48550/arXiv.2406.13232. URL https://arxiv.org/abs/2406.13232.

Alexander M. Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vázquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Neeraj Varshney, Satyam Raj, Venkatesh Mishra, Agneet Chatterjee, Ritika Sarkar, Amir Saeidi, and Chitta Baral. Investigating and addressing hallucinations of llms in tasks involving negation. In *Proceedings of the 5th Workshop on Trustworthy Natural Language Processing (TrustNLP 2025)*. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.trustnlp-main.37/.

Jiuding Yang, Weidong Guo, Kaitong Yang, Xiangyang Li, Zhuwei Rao, Yu Xu, and Di Niu. Optimizing and testing instruction-following: Analyzing the impact of fine-grained instruction variants on instruction-tuned llms. *arXiv preprint arXiv:2406.11301*, 2024. doi: 10.48550/arXiv.2406.11301. URL https://arxiv.org/abs/2406.11301.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. Crepe: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10457–10480. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.583. URL https://aclanthology.org/2023.acl-long.583/.

Huaiwen Zhang, Yu Chen, Ming Wang, and Shi Feng. Feel: A framework for evaluating emotional support capability with large language models. In De-Shuang Huang, Wei Chen, and Qinhu Zhang (eds.), *Advanced Intelligent Computing Technology and Applications*, pp. 96–107. Springer Nature Singapore, Singapore, 2024. ISBN 978-981-97-5618-6.

Kun Zhao, Bohao Yang, Chen Tang, Chenghua Lin, and Liang Zhan. Slide: A framework integrating small and large language models for open-domain dialogues evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.911. URL https://aclanthology.org/2024.findings-acl.911.pdf.

Irune Zubiaga, Aitor Soroa, and Rodrigo Agerri. A llm-based ranking method for the evaluation of automatic counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.559. URL https://aclanthology.org/2024.findings-emnlp.559/.