Argumentation Based Systems Journal

www.absj.org
eISSN: 3120-1024

Reasoning About the Non-Existent in Large Language Models: Benchmark Success, Negation Failure, and Exploratory Creativity

Abdullah ÖNDEN[1*], İsmail ÖNDEN[2]

[1] Department of Computer Engineering, Faculty of Computer and Information Technologies, Istanbul University, Istanbul, Türkiye
[2] Department of AI and Data Engineering, Faculty of Computer and Information Technologies, Istanbul University, Istanbul, Türkiye

ARTICLE INFO

Article history:
Received 2 February 2026
Received in revised form 18 March 20226
Accepted 24 March 2026
Available online 27 March 2026

Keywords:
Frontier Large Language Models; Counterfactual Reasoning; Negation Blindness; Simulation-Reality Gap; Benchmark Saturation; CRASS; Live Evaluation

ABSTRACT

Can artificial intelligence reason about things that do not exist, are not true, or are merely hypothetical? This paper addresses these questions with a two-part analysis of non-existence reasoning in state-of-the-art large language models. The analysis includes four components: counterfactual inference, negation handling, creative novelty, and multi-turn persistence of imaginary knowledge. Part 1 uses a benchmark-calibrated forecasting model to predict future performance for 12 models on 15,300 simulated datapoints. Part 2 revisits the question with live API evaluations of 5 state-of-the-art models on over 654 items covering CRASS-style counterfactual questions, negation tasks, creativity prompts, and multi-turn dialogues about imaginary topics. The outcome is a differentiated capability profile rather than a simple yes or no answer to the question of whether AI can reason about non-existence. The sampled CRASS items produced near-ceiling performance relative to the published human baseline, suggesting that some portions of the benchmark may no longer provide strong discriminative power among state-of-the-art models. By contrast, negation was associated with statistically significant performance drops in the four complete model runs, with the same directional tendency also visible in the partial fifth run. The conclusions about creative novelty and multi-turn imaginary persistence are more provisional since they rested on proxy measures and heuristic contradiction detection. In substance, the results imply that top models can execute certain types of structured counterfactual reasoning effectively but display a robust and replicable deficit with respect to negation; indications of difficulty with creative novelty and multi-turn imaginative persistence are present but rest on weaker proxy measures and should be treated as provisional. In method, they underscore the importance of treating benchmark-calibrated forecasts as perishable hypotheses that must be revalidated with live model evaluations in a rapidly evolving LLM environment.

1. Introduction

Counterfactual scenarios, absent concepts, negated propositions, and imaginary constructs all involve situations that do not exist. How well can artificial intelligence reason about them? This question, which traces back to Turing's (1950) [1] inquiry into machine thinking, is challenging for several reasons. While large language models (LLMs) have achieved impressive results in many areas

* *Corresponding author.*
E-mail address: abdullah.onden@istanbul.edu.tr

https://doi.org/10.59543/avhexs22

[2, 3, 4], the ability to represent what is not real plausibly demarcates the boundary between powerful pattern recognition and deeper forms of understanding [5, 6]. The question is of practical interest as well. Counterfactuals underlie causal explanation, planning, diagnosis, and moral judgment, while negation and absence underlie exclusion, constraint handling, and safety-critical reasoning [7, 8, 9]. A system that writes fluently but fails at what did not happen, what is ruled out, or what has never been, may still be brittle precisely in the conditions where users most trust fluent language.

Recent model progress makes the issue harder to study than one might infer from looking at older benchmark results. Top-line scores on broad benchmarks have been interpreted as evidence of emerging reasoning abilities [2, 3, 4, 10, 11]. However, these scores might reflect several different abilities at once: retrieval of familiar patterns, recombination of learned patterns, and genuine manipulation of states of affairs that are absent from the world. A direct empirical comparison requires an evaluation design that can separate capability from benchmark aging. With this need in mind, we adopt a two-phase design. In Phase 1, we built a simulation framework that was calibrated to published benchmark results to generate structured forecasts of LLM performance across four hypotheses. In Phase 2, we subjected a smaller set of frontier models to direct API evaluation on matched task families. The design serves two purposes: (1) to probe substantive questions about reasoning with non-existence and (2) to evaluate how well literature-calibrated forecasts travel to a rapidly changing model landscape.

The results suggest that the answer to the question above is nuanced rather than simple. On the one hand, simulation-based predictions, even when calibrated to peer-reviewed benchmarks, significantly underestimated current frontier performance on the sampled CRASS items. This suggests that at least some counterfactual probes have lost discriminative power. On the other hand, the improvement is not uniform across all forms of non-existence reasoning. On the same set of models, negation still produced large and consistent performance drops, while the evidence for creative novelty and multi-turn imaginary consistency was still exploratory in that those constructs were measured with weaker proxies.

1.1. Research Questions

RQ1: To what extent can current LLMs reason about counterfactual scenarios, and has performance changed since earlier benchmark studies?

RQ2: Do LLMs exhibit systematic performance degradation when reasoning involves negation and absence?

RQ3: How do LLMs perform on creative tasks requiring the generation of genuinely non-existent concepts?

RQ4: Can LLMs maintain coherent representations of imaginary constructs across multi-turn dialogues?

RQ5: How accurately do simulation-based predictions estimate real LLM performance?

Accordingly, the paper makes three contributions. First, it brings counterfactual reasoning, negation, creative generation, and imaginary consistency into a single analytical frame rather than treating them as separate deficits. Second, it introduces a two-phase design that compares benchmark-calibrated forecasts with live API measurements, allowing the study to speak both to substantive capability and to evaluation drift. This leads to the second main conclusion of the paper: claims about LLM reasoning about non-existence should be grounded in a more specific and temporally-limited interpretation. Some types of benchmarked non-existence reasoning may be undergoing rapid improvement, while linguistically-mediated absence reasoning appears to represent a more persistent vulnerability.

2. Theoretical Background

2.1. Non-Existence Reasoning

Non-existence reasoning is a long-standing philosophical topic, stretching back to Parmenides and forward through Sartre's phenomenology [12], Heidegger's [13] confrontation with nothingness, and possible world semantics [14, 15] to Pearl's causal hierarchy [16]. Counterfactual reasoning is the top rung of the ladder and is a key component of causal reasoning [17, 7, 8]. Frohberg and Binder [18] introduced the CRASS benchmark which contains 274 counterfactual conditionals and a human baseline of 95.6. Chen et al. [19] introduced the CounterBench benchmark finding near-chance performance on formal counterfactual inference. Chi et al. [20] introduced the CausalProbe-2024 benchmark which shows that LLMs are functioning at associational levels, consistent with evidence that LLMs may produce hallucinated causal claims [21]. García-Ferrero et al. [22] show using 400,000 sentences that LLMs struggle with negation even when fine-tuned. Alhamoud et al. [23] further demonstrate that vision-language models also fail to process negation correctly. Truong et al. [24] and Anschütz et al. [25] confirmed that LLMs consistently fail to capture the lexical semantics of negation across model sizes, while Kassner and Schütze [26] showed that negated probes retrieve the same answers as affirmative ones. Mondorf and Plank [27] show that proficient models struggle with suppositional reasoning. As a result of benchmark saturation [28, 29] where models reach near-ceiling performance on what were challenging benchmarks, it can be increasingly difficult to find benchmarks that measure the desired capability. In the fast-moving LLM regime, a benchmark may still measure something, but it may not measure the contrast that researchers think it does.

2.2. Counterfactual Benchmarks

Existing counterfactual benchmarks differ in whether they target formal intervention-style reasoning, everyday alternative scenarios, or causally framed question answering. This matters because strong performance on one family does not guarantee transfer to another, particularly when the benchmark has aged relative to the frontier model class [18, 19, 20].

2.3. Negation Understanding

Negation is not a niche linguistic phenomenon but a core mechanism for expressing exclusion, contradiction, and absence. Prior work suggests that LLMs often preserve the topical content of a sentence while failing to reliably invert its truth conditions, which makes negation a particularly stringent test of non-existence reasoning [26, 25, 24, 23].

2.4. Benchmark Saturation

Benchmark saturation occurs when a once-discriminative dataset no longer separates frontier systems in a meaningful way. In that setting, historical scores remain informative about past capability levels but become weaker evidence about present limitations, which is why live revalidation is necessary [28, 29].

Taken together, these various literatures suggest that non-existence reasoning should not be understood as a yes-or-no capability. A model may answer standard counterfactual questions correctly, but struggle when the absence is signaled in a more linguistically mediated way, or when a novel example must be discriminated from examples that have come before, or when an imaginary entity must be maintained across conversational turns. This is part of the motivation for the two-phase design of the current study: it tests a series of related non-existence reasoning capabilities, and it also tests the rate at which benchmark-based inferences can become outdated. The simulation (Phase 1) is used to generate predictions, and the empirical study (Phase 2) is used to test those predictions. This serves a dual purpose: it tests the substantive hypotheses, but it also tests the validity of the simulation itself. Because H3 and H4 rely on more ad hoc measurement procedures

than H1 and H2, those results are understood to be more provisional and are interpreted more cautiously and in an exploratory manner where necessary.

3. Methodology

3.1. Two-Phase Design

The study therefore used a two-phase design. Phase 1 translated prior benchmark evidence into forward-looking performance expectations, whereas Phase 2 subjected those expectations to live API testing under a common prompting and evaluation protocol.

3.2. Phase 1: Simulation (12 Models, 15,300 Data Points)

The simulation performance parameters were tuned to match the performance of models on the CRASS benchmark [18], CounterBench [19], and the NegationBench [22]. The models simulated were: GPT-4o, GPT-4o-mini, Claude 3.5 Sonnet, Claude 3 Haiku, Gemini 1.5 Pro, Gemini 1.5 Flash, Llama 3.1 70B, Llama 3.1 8B, Mistral Large, Mixtral 8x22B, Qwen 2.5 72B, DeepSeek-V3. Phase 1 should be understood as a benchmark-calibrated prediction exercise rather than as a mechanistic cognitive simulation: it predicts task performance using published benchmarks, but it does not directly simulate the model's performance on new prompts.

3.3. Phase 2: Real API Evaluation (5 Models, 654+ Items)

Five state-of-the-art models were targeted for live evaluation: GPT-4o, GPT-4o-mini, Claude 3.5 Sonnet, Gemini 1.5 Pro, Gemini 1.5 Flash. All tasks used temperature = 0.0 in order to minimize decoding noise, following standard practice in causal reasoning evaluation [30]. Due to API instability, however, not every model yielded fully usable outputs for every task; in H2, inferential statistics are therefore reported for the four complete runs, while the partial Gemini 1.5 Pro run is shown descriptively only. For H1, the comparison to the published human baseline should be understood as descriptive: no contemporaneous human control group was collected under the same protocol.

H1: 44 questions from the CRASS benchmark (BIG-Bench version). Published human baseline: 95.6.

H2: 300 sentences from HiTZ/This-is-not-a-dataset (EMNLP 2023). Three conditions: verbal, non-verbal, affirmative control.

H3: 50 prompts for creative generation of non-existent objects across 5 categories.

H4: 30 multi-turn scenarios (4 turns each) across 6 categories testing imaginary construct consistency. We employed the Chi-square test and Fisher's Exact test for proportions, Cohen's h for the difference between proportions, and the Kruskal-Wallis H-test for continuous variables. For all statistical tests, we adopted an alpha level of 0.05. Effect sizes were interpreted following Cohen's [31] conventions. For comparison with the literature-based human baseline, the results are descriptive as no concurrently sampled human population was available.

H1 Prediction: AI 58.1% vs. human 95.6% (gap: 37.5 percentage points; very large predicted standardized difference). H2 Prediction: Average accuracy decrease for negated sentences: 18.9%. H3 Prediction: Average training similarity: 0.670; $\rho = -0.78$ with novelty. H4 Prediction: 7 to 10/12 models will show a significant turn-contradiction correlation.

3.4. Statistical Analysis

The confirmatory analyses for H1 and H2 centered on accuracy differences and effect sizes, whereas H3 and H4 were treated as exploratory because the available measures captured diversity and explicit contradiction signals more directly than ground-truth novelty or latent inconsistency. Accordingly, inferential claims are strongest for H1–H2 and more provisional for H3–H4.

4. Results

4.1. Phase 1

Phase 1 produced a pessimistic forecast for counterfactual performance and anticipated a persistent negation penalty across models. It also suggested limited novelty and declining coherence across longer imaginary interactions, thereby generating clear hypotheses for the live evaluation stage.

4.2. Phase 2: Real Results

4.2.1. H1: Counterfactual Reasoning — Prediction Overturned

Table 1 presents the Phase 2 CRASS results (44 questions, real API data). Note: Human baseline = 95.6% [18]. Inferential statistics are not reported for this comparison: the human baseline is literature-derived and no contemporaneous matched human sample was collected under the same protocol.

On the 44-question subset sampled in Figure 1, all models performed within a narrow band of high accuracy centered around the published human baseline.

Table 1. Phase 2: CRASS Counterfactual Reasoning (44 questions, real API data)

Model	n	Accuracy	Gap
GPT-4o	44	95.45%	+0.15%
GPT-4o-mini	44	90.91%	−4.69%
Claude 3.5 Sonnet	44	95.45%	+0.15%
Gemini 1.5 Pro	44	97.73%	+2.13%
Gemini 1.5 Flash	44	97.73%	+2.13%
Mean	220	95.45%	−0.15%

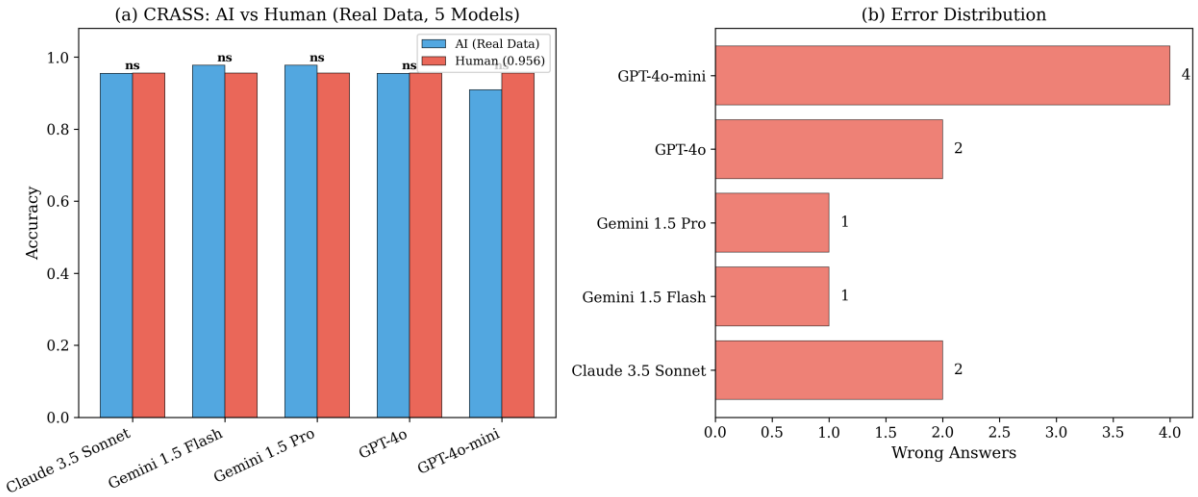

Figure 1. CRASS Counterfactual Reasoning: AI vs. published human baseline

The mean AI accuracy on this task was 95.45%, very similar to the published human baseline of 95.6%. Gemini 1.5 Pro and Gemini 1.5 Flash both scored 97.73% (43/44) on the subset sampled here. This is in stark contrast to the Phase 1 prediction of 58.1% and suggests that the benchmark-calibrated prediction underestimated the current state-of-the-art performance on this task. Because the human baseline was obtained from literature rather than sampled anew here, this comparison should be understood as descriptive evidence of benchmark aging on the sampled subset, not as a formal human-versus-model parity claim.

4.2.2. H2: Negation Understanding — Prediction Confirmed

Table 2 presents the Phase 2 Negation Understanding results (300 items, real API data). Inferential statistics are reported for the four complete model runs; the partial Gemini 1.5 Pro run is shown descriptively in Figure 2 only. Kruskal-Wallis across negation types: $H = 50.24$, $p < 0.001$.

Table 2. Phase 2: Negation Understanding (300 items, real API data).

Model	Affirm.	Negated	Drop	$\chi^2(1)$	p-value	Cohen's h
GPT-4o	88.6%	72.3%	18.4%	9.59	0.002**	0.419
GPT-4o-mini	81.0%	67.7%	16.4%	5.35	0.021*	0.306
Claude 3.5 Sonnet	84.8%	56.1%	33.8%	21.91	<0.001***	0.647
Gemini 1.5 Flash	80.0%	62.9%	21.4%	5.35	0.021*	0.382

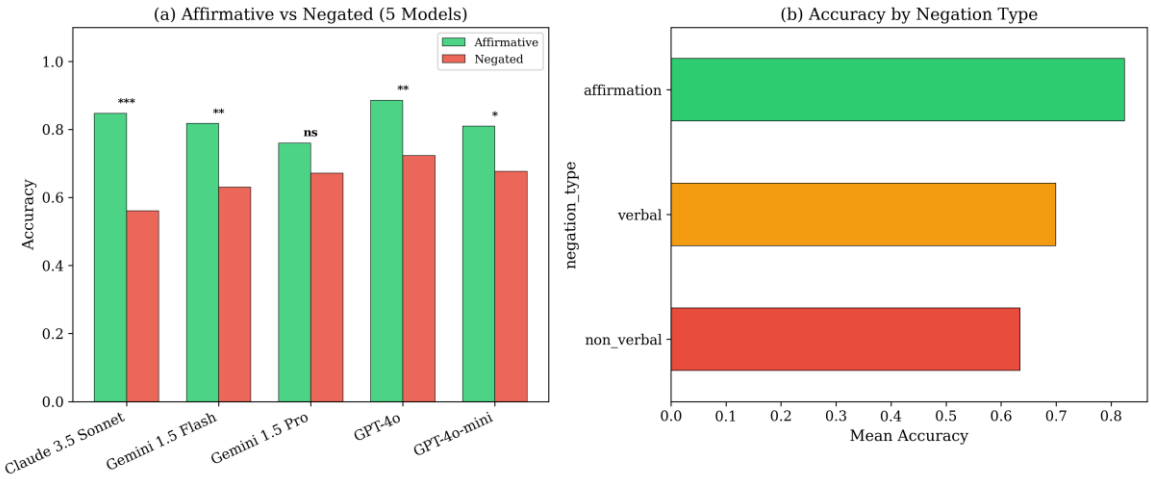

Figure 2. Negation Understanding: Affirmative vs. Negated accuracy (real data)

Four of the five model runs showed statistically significant decreases in accuracy for negated sentences. The partial Gemini 1.5 Pro run showed the same directional pattern, but it is treated descriptively rather than inferentially because the run was incomplete. The largest decrease among the complete runs was observed for Claude 3.5 Sonnet (33.8%, chi-square = 21.91, $p < 0.001$, Cohen's $h = 0.647$). Non-verbal negation was consistently the most difficult negation subtype (mean accuracy = 63.2%), followed by verbal negation (mean accuracy = 77.8%), whereas affirmative controls were generally high (mean accuracy = 87.9%). Unlike H1, this finding aligns directionally with the Phase 1 prediction and thus represents the strongest evidence for a persistent, task-specific deficiency in this study.

4.2.3. H3: Creative Novelty (Exploratory)

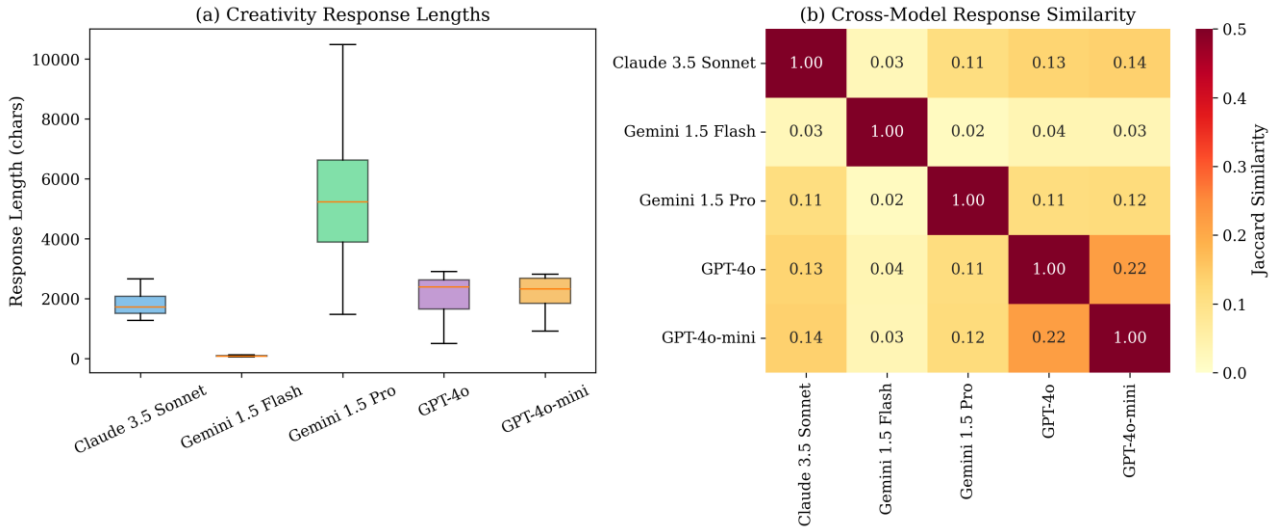

Figure 3. Creativity analysis: (a) Distribution of response length by model; (b) Pairwise Jaccard similarity between word sets

Gemini 1.5 Pro generated by far the longest average responses (mean: 5,224 characters), followed by GPT-4o-mini (mean: 2,172), GPT-4o (mean: 2,117), and Claude 3.5 Sonnet (mean: 1,797). Gemini 1.5 Flash generated substantially shorter responses (mean: 81 characters). The mean Jaccard similarity was only 0.059; this suggests the models are using a variety of words and concepts, but this metric reflects output variability rather than confirmed novelty, so H3 should be viewed as exploratory evidence of response diversity rather than confirmatory evidence of creative novelty.

4.2.4. H4: Imaginary Construct Consistency (Exploratory)

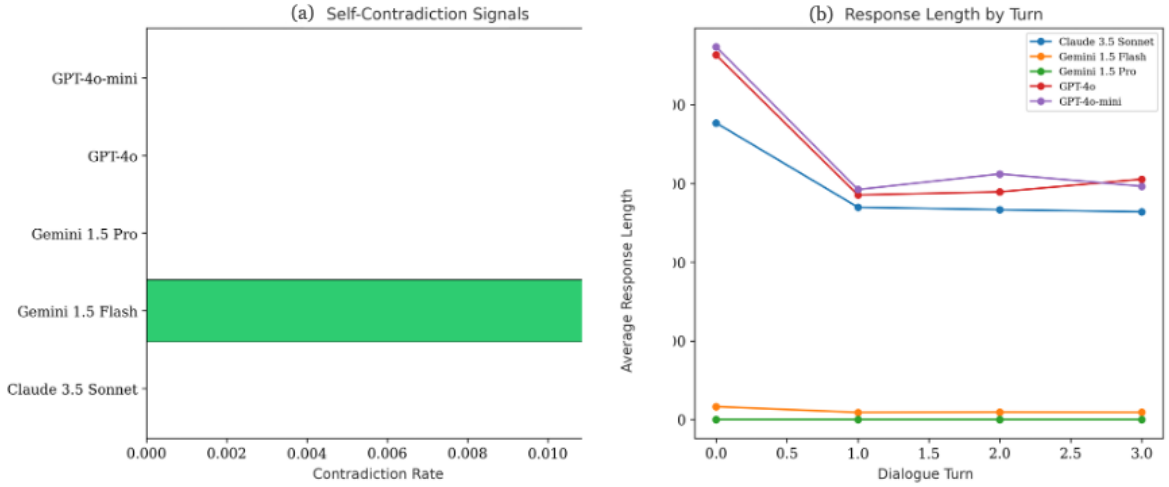

Figure 4. Consistency analysis: (a) Number of self-contradiction signals per model; (b) Average response length across turns

Heuristic contradiction detection yielded a rate below 1.1 for each model. GPT-4o, GPT-4o-mini, and Claude 3.5 Sonnet also sustained average response lengths of about 1,500–1,700 characters across turns. The models were therefore engaged with the simulated worlds, but the contradiction detector was designed to be conservative and so this analysis likely missed many subtle changes in world-state, feature drift, and latent contradiction. Thus, H4 should also be viewed as exploratory.

4.3. The Simulation–Reality Gap

Table 3 presents the Simulation-Reality Gap: Phase 1 Predictions vs. Phase 2 Results. The pattern is clear: H1 (Counterfactual) was overturned (AI predicted 58.1%, actual 95.45%), indicating benchmark aging on the sampled subset. H2 (Negation drop) was largely confirmed (predicted 18.9%, actual 22.5%), with 4/5 complete runs showing significant drops. H3 (Creativity) was partial, suggesting more analysis is needed. H4 (Consistency) was inconclusive, with better methods needed for future research.

Table 3. Simulation-Reality Gap: Phase 1 Predictions vs. Phase 2 Results

Hypothesis	Phase 1	Phase 2	Match?	Implication
H1: Counterfactual	AI: 58.1%	AI: 95.45%	Overturned	Evidence of benchmark aging on the sampled subset
H2: Negation drop	18.9% drop	22.5% drop	Largely Confirmed	4/5 significant; persistent
H3: Creativity	Low novelty	Divergent outputs	Partial	More analysis needed
H4: Consistency	Increasing errors	Low explicit errors	Inconclusive	Better methods needed

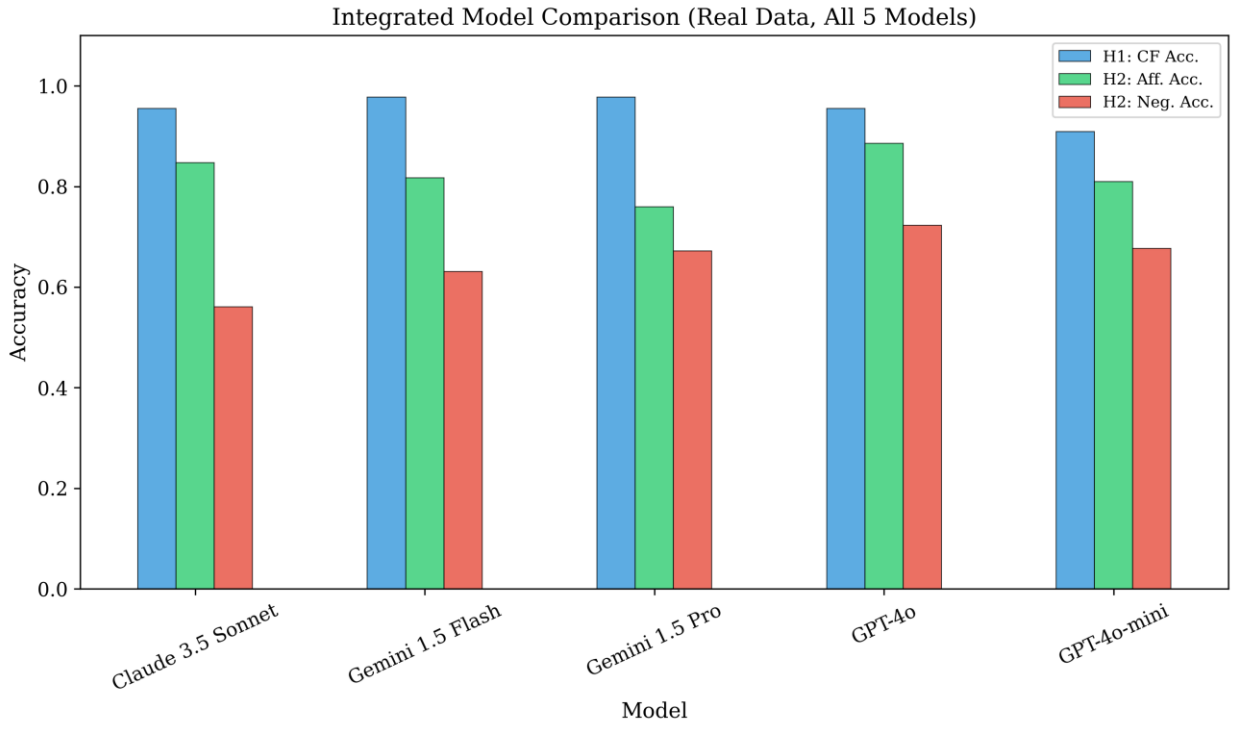

Figure 5. Combined model overview for all hypotheses

5. Discussion

5.1. Interpreting the Main Result

The most important takeaway is that these results do not warrant the conclusion that current LLMs either can or cannot reason about the non-existent. The frontier models performed within a narrow band of high accuracy on the sampled CRASS items, descriptively close to the published human baseline. This matters, because it means at least one obvious class of counterfactual questions is no longer a clean signal of the difference between today's systems and the ones that came before.

However, these results also should not be interpreted to mean that the current generation of LLMs has mastered this ability. The CRASS analysis is based on a subset of the test, not the full one, and the human comparison is to a published baseline rather than a newly collected matched sample. Furthermore, the overall picture is more nuanced: the frontier models are strong on H1 and on structured counterfactual questions, but display a large drop in performance on H2, and only exploratory evidence for creative novelty and multi-turn representational consistency. The more defensible conclusion is therefore narrower, but more sustainable: current LLMs appear very strong on some benchmarked forms of non-existence reasoning, but are far less uniformly strong once the task requires semantic exclusion, open-ended novelty, or multi-turn representational consistency [32].

5.2. Benchmark Saturation and Temporal Validity

One of the most direct takeaways of this paper is that benchmark results have an expiration date. The Phase 1 analysis predicted that we would observe a sizeable human-model performance gap, because that analysis was grounded in prior published counterfactual results. Instead, in Phase 2, we observed near-ceiling performance on the sampled CRASS items. This is itself a contribution; it shows how the empirical meaning of a benchmark can change, as model quality, prompting strategies, and the surrounding training ecosystem evolve more rapidly than evaluation frameworks do [33, 34].

This observation has implications that go beyond the present study. If an old benchmark is continually cited as evidence of a purported limitation, it is easy for the community to conflate past performance with present abilities. On the flip side, high absolute performance on an aged benchmark may also engender a false sense of security about reasoning abilities. Both mistakes stem from the same underlying issue: claims about ability become decoupled from fresh empirical measurement. For rapidly improving models, we recommend treating the validity of the benchmark as an empirical question to be tested, rather than an assumption to be made.

5.3. Negation as a Persistent Bottleneck

Negation is the strongest weakness we detected in the study. Across the four complete model runs—and descriptively in the partial fifth run—negated items resulted in marked performance drops relative to affirmative controls. This should not be dismissed as a small linguistic quirk. Negation is one of the principal ways by which natural language encodes exclusion, exception, absence, and contradiction.

The pattern is also theoretically illuminating. Models performed better when counterfactual content was embedded in familiar conditional structures than when absence was signaled through explicit negation. This asymmetry implies that the surface form of negation, rather than the abstract content of non-existence, is the proximate cause of the difficulty. This would be consistent with evidence that negation tokens may have limited effect on internal representations and that models may process negated sentences by partially reversing affirmative representations rather than by building up a separate semantic representation.

The practical upshot is that improvements on structured counterfactual tests do not automatically generalize to tasks that rely on syntactic negation to signal exclusion or absence. This practically matters because many real tasks are framed in terms of exclusion: what a diagnosis rules out, what a contract prohibits, what a safety system must not allow. In such settings, a model that fluently reasons about scenarios yet systematically misunderstands negated constraints is a reliability risk even if it scores well on standard benchmarks. Our results suggest that negation-specific testing should be a standard part of deployment evaluations rather than an afterthought.

5.4. Theoretical Implications

Our results support a differential rather than a monolithic account of non-existent reasoning. Rather than asking whether a model can or cannot reason about the non-existent, it is more accurate to ask which types of non-existent reasoning a model can handle and which it cannot. On our evidence: structured counterfactual questions of the CRASS type appear to be within the competence of current frontier models, while syntactically-explicit negation consistently produces degradation, and open-ended non-existent tasks (novelty, imaginative consistency) remain difficult to evaluate decisively.

Relative to skeptical critiques like the stochastic parrots view [35] and Searle's [36] Chinese Room argument, our results are more nuanced than a simple Yes or No answer to the question of whether LLMs are capable of cognition. The near-ceiling counterfactual performance is hard to reconcile with a pure retrieval account unless CRASS items are well-represented in training data. Notably, Si et al. [37] found that LLM-generated research ideas can be rated as more novel than human expert ideas, complicating simple accounts of LLMs as mere pattern matchers. At the same time, the negation results imply that structural comprehension of absence is brittle, which is in line with the surface-form processing concerns. The intellectually honest thing to say would seem to be that different types of non-existent reasoning make different demands, and current models meet those demands to different extents. A model that answers familiar counterfactual questions at accuracy levels descriptively close to the published human baseline while showing large negation deficits is neither straightforwardly cognitively capable nor straightforwardly merely pattern-matching: it is managing both, in different ways, depending on the task.

We believe that Pearl's causal hierarchy remains a valuable framework for understanding this phenomenon, but our paper does not purport to settle where LLMs fall on that hierarchy. Our counterfactual data is compatible with some level of competence at the interventional or counterfactual level, but our tasks were not engineered to distinguish causal inference from conditional language modeling. Doing so would have required tasks that specifically disentangled statistical association from causal structure, and that remains a promising direction for future research.

5.5. Practical Implications

For practical deployment, the key takeaway is not a blanket message of "do trust LLMs with non-existence reasoning" or "don't trust LLMs with non-existence reasoning." Rather, it is to draw distinctions between task types. Applications that demand structured counterfactual reasoning of the sort captured by the CRASS inventory can make use of current frontier performance in this area. Applications that rely on correct negation processing, particularly multi-negation or negation in safety-critical constraints, should be approached with a far more cautious attitude and should include negation-specific test sets before deployment.

This distinction is particularly relevant for professional domains. In legal reasoning, the distinction between an obligation and its negation is a matter of logical life and death. In medical decision support, what a test rules out may be as crucial as what it confirms. In automated content

moderation, correctly identifying what is prohibited requires accurately processing negation. Our results indicate that current systems may be able to handle the affirmative version of these tasks better than the negated version, and that is precisely the asymmetry that matters for high-stakes applications.

Our exploratory results on novelty and imaginary consistency also warrant caution in applications where novel concept generation or sustained world-building is important. True creativity requires conceiving of what does not yet exist [38, 39], and computational creativity research has distinguished exploratory from transformational creativity [40]. Recent embedding analyses suggest that LLM outputs remain within the statistical neighborhood of training data even when rated as creative by human evaluators [41], and LLM-generated content may reduce collective diversity in creative ideation [42, 43]. The proxy metrics we used were not up to the task of determining whether models generate genuinely novel outputs or whether they maintain imaginative consistency over extended interactions. Applications in creative assistance, game design, and interactive fiction that depend on these capacities should be evaluated using more targeted instruments than response length or heuristic contradiction detection.

5.6. Methodological Implications

Methodologically, the paper argues for a more rigorous separation between benchmark-calibrated forecasting and live model evaluation. Phase 1 produced seemingly reasonable predictions from a principled framework, yet those predictions were badly off the mark for H1. The discrepancy is not a function of poor calibration but of the underlying problem: the validity of a benchmark is a moving target in a field where training data, fine-tuning objectives, and model scale are all evolving at a rapid clip. This means that predictions derived from peer-reviewed results should be treated as perishable hypotheses rather than durable estimates.

The general moral here is that LLM evaluation requires repeated revalidation rather than periodic benchmarking. As the simulation-reality gap in this study makes clear, a prediction that was not only plausible but peer-reviewed a few years ago can be overturned by today's frontier models. The field would benefit from evaluation infrastructure that tracked not only model performance but also benchmark drift: the extent to which a test is still a measure of the capacity it was designed to probe. In the absence of such tracking, the gap between nominal and actual capability may grow without our even noticing.

5.7. Limitations and Future Research

Several caveats must be attached to these conclusions. The second phase used five target models rather than the twelve simulated, and not every target model yielded usable outputs for every task because of API instability. The CRASS result is based on 44 of 274 questions, and therefore should be understood as about the subset, not the entire dataset. The human comparison uses a pre-existing baseline, rather than a novel matched sample, which restricts our ability to make direct inferences between humans and models. H3 uses proxy measures of diversity rather than third-party novelty ratings or information-retrieval-based overlap metrics, and H4 relies on heuristic contradiction detection that probably fails to catch more subtle contradictions. Future work should consider NLI-based contradiction detection methods [44], self-consistency approaches for hallucination detection [45], and multi-session dialogue evaluation frameworks [46].

These caveats also have direct implications for future research. The most valuable follow-ups would include live replication on the full counterfactual battery, contemporaneous human comparison on matched prompts, stronger novelty assessment using information retrieval or blinded human ratings, and more robust contradiction metrics using NLI-based or hand-coded evaluations. More generally, future work should prioritize novel test items so that claims about reasoning

capabilities or deficits can be grounded in the current state of the model landscape rather than in the historical age of the benchmark.

6. Conclusion

By combining benchmark-calibrated simulation with live API evaluation in a two-phase study, this research provides a more robust understanding of non-existence reasoning in modern LLMs than could be provided by either method in isolation.

The results support four primary conclusions. First, on the subset of CRASS items tested, state-of-the-art models now performed within a narrow band of high accuracy on the sampled subset, descriptively close to the published human baseline, consistent with benchmark aging and reduced discriminative power on this variety of counterfactual questions. Second, negation is the most consistent and replicable deficit: all four complete model runs showed significant performance declines when test items were framed in their negated form, while the partial Gemini 1.5 Pro run showed the same directional tendency descriptively. Third, benchmark-based simulation may drastically underestimate current abilities when it attempts to extrapolate from frozen historical data to a fast-moving model landscape, echoing recent findings that causal reasoning evaluations require continual updating [47]. Fourth, any claims about the novelty of creatively generated content or the coherence of reasoning about imaginary scenarios should be accepted as provisional pending the development of better metrics.

The overall implication of this is that we should not treat non-existence reasoning as monolithic. While the most advanced LLMs appear to handle some forms of structured counterfactual reasoning well, they display a consistent and replicable deficit when non-existence is expressed through syntactic negation or exclusion, and show more preliminary—and more provisional—evidence of difficulty with open-ended creative generation and multi-turn imaginative consistency. For researchers, this means that we require more fine-grained constructs, fresher benchmarks, and repeated real-world validation. For practitioners, this means that excellent performance on one type of reasoning benchmark should not be taken to indicate overall trustworthiness on all types of hypothetical, absent, or imaginary reasoning. The core takeaway of the paper is thus inherently double-edged: the most recent LLMs are more capable than the old benchmarks indicate, but they are also more spotty than suggested by aggregate summaries of performance.

Author Contributions

Conceptualization, İ.Ö - A.Ö.; methodology, İ.Ö - A.Ö.; software, A.Ö.; validation, İ.Ö - A.Ö.; formal analysis, A.Ö.; investigation, A.Ö.; resources, A.Ö.; data curation, A.Ö.; writing—original draft preparation, İ.Ö - A.Ö.; writing—review and editing, İ.Ö - A.Ö.; visualization, A.Ö. The authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

Data supporting the reported results, including evaluation prompts and model outputs, are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was not funded by any grant.

References

[1] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433
[2] OpenAI. (2024). *GPT-4o system card*. Technical report. https://openai.com/index/gpt-4o-system-card
[3] Anthropic. (2024). *Claude 3.5 Sonnet model card addendum*. Technical report. https://www.anthropic.com/news/claude-3-5-sonnet
[4] Google DeepMind. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*. https://doi.org/10.48550/arXiv.2403.05530
[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762
[6] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903
[7] Byrne, R. M. J. (2005). *The rational imagination*. MIT Press. https://doi.org/10.7551/mitpress/5756.001.0001
[8] Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, 12(2), 168–192. https://doi.org/10.1177/1088868308316091
[9] Pearl, J., & Mackenzie, D. (2018). *The book of why*. Basic Books.
[10] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*. https://doi.org/10.48550/arXiv.2303.12712
[11] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. https://doi.org/10.48550/arXiv.2206.04615
[12] Sartre, J.-P. (1943). *L'Être et le néant: Essai d'ontologie phénoménologique*. Gallimard.
[13] Heidegger, M. (1929). *Was ist Metaphysik?* Friedrich Cohen.
[14] Lewis, D. (1973). *Counterfactuals*. Blackwell.
[15] Kripke, S. A. (1980). *Naming and necessity*. Harvard University Press.
[16] Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161
[17] Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, 121(1), 133–148. https://doi.org/10.1037/0033-2909.121.1.133
[18] Frohberg, J., & Binder, F. (2022). CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 2126–2140). https://doi.org/10.48550/arXiv.2112.11941
[19] Chen, Y., Kotonya, N., & Toni, F. (2025). CounterBench: Exploring counterfactuals in LLMs through evaluation and training. *arXiv preprint arXiv:2502.11008*. https://doi.org/10.48550/arXiv.2502.11008
[20] Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., Liu, T., & Han, B. (2024). Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, 37, 96640–96670. https://doi.org/10.48550/arXiv.2410.21597
[21] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models. *arXiv preprint arXiv:2311.05232*. https://doi.org/10.48550/arXiv.2311.05232
[22] García-Ferrero, I., Altuna, B., Alvez, J., Gonzalez-Dios, I., & Rigau, G. (2023). This is not a dataset: A large negation benchmark to challenge large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 8596–8615). https://doi.org/10.18653/v1/2023.emnlp-main.531
[23] Alhamoud, K., Alshammari, S., Tian, Y., Li, G., Torr, P. H. S., Kim, Y., & Ghassemi, M. (2025). Vision-language models do not understand negation. *arXiv preprint*. https://doi.org/10.48550/arXiv.2501.09425
[24] Truong, T. H., Baldwin, T., Verspoor, K., & Cohn, T. (2023). Language models are not naysayers: An analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*. https://doi.org/10.18653/v1/2023.starsem-1.10
[25] Anschütz, M., Jalota, D., Mondorf, P., & Plank, B. (2023). On the language understanding capabilities of ChatGPT and GPT-4 with respect to negation. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics* (pp. 284–295). https://doi.org/10.18653/v1/2023.starsem-1.29

[26] Kassner, N., & Schütze, H. (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* (pp. 7811–7818). https://doi.org/10.18653/v1/2020.acl-main.698

[27] Mondorf, P., & Plank, B. (2024). Liar, liar, logical mire: A benchmark for suppositional reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 7114–7137). https://doi.org/10.18653/v1/2024.emnlp-main.404

[28] Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., & Williams, A. (2021). Dynabench: Rethinking benchmarking in NLP. In *Proceedings of NAACL 2021*. https://doi.org/10.18653/v1/2021.naacl-main.324

[29] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations 2021*. https://doi.org/10.48550/arXiv.2009.03300

[30] Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez Adauto, F., Kleiman-Weiner, M., Sachan, M., & Schölkopf, B. (2023). CLadder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems (NeurIPS 2023)*. https://doi.org/10.48550/arXiv.2312.04350

[31] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

[32] Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Le Bras, R., Hwang, J. D., Sanyal, S., Ren, X., Ettinger, A., Harchaoui, Z., & Choi, Y. (2023). Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36. https://doi.org/10.48550/arXiv.2305.18654

[33] Ott, S., Heuss, M., & Markert, K. (2022). Mapping the landscape of evaluation methodology for NLP systems. In *Findings of EMNLP 2022*. https://doi.org/10.48550/arXiv.2212.04329

[34] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., & Hudson, D. A. (2023). Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1), 140–146. https://doi.org/10.1111/nyas.15007

[35] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

[36] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. https://doi.org/10.1017/S0140525X00005756

[37] Si, C., Yang, D., & Hashimoto, T. (2024). Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv preprint arXiv:2409.04109*. https://doi.org/10.48550/arXiv.2409.04109

[38] Guilford, J. P. (1950). Creativity. *American Psychologist*, 5(9), 444–454. https://doi.org/10.1037/h0063487

[39] Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd ed.). Routledge.

[40] Colton, S., Pease, A., & Charnley, J. (2012). Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the Third International Conference on Computational Creativity*. https://research.gold.ac.uk/id/eprint/7670/

[41] Franceschelli, G., & Musolesi, M. (2023). On the creativity of large language models. *arXiv preprint arXiv:2304.00008*. https://doi.org/10.48550/arXiv.2304.00008

[42] Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. In *Proceedings of the 15th International Conference on Computational Creativity*. https://doi.org/10.48550/arXiv.2402.01536

[43] Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022). Putting GPT-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity* (pp. 164–168). https://doi.org/10.48550/arXiv.2206.08932

[44] Nguyen, V. B., Youssef, P., Seifert, C., & Schlötterer, J. (2024). LLMs for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 14809–14824). https://doi.org/10.18653/v1/2024.findings-emnlp.870

[45] Manakul, P., Liusie, A., & Gales, M. J. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP 2023* (pp. 9004–9017). https://doi.org/10.18653/v1/2023.emnlp-main.557

[46] Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., & Fang, Y. (2024). Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2024.acl-long.747

[47] Willig, M., Zecevic, M., von Kügelgen, J., & Kersting, K. (2022). Can foundation models talk causality? *arXiv preprint arXiv:2206.10591*. https://doi.org/10.48550/arXiv.2206.10591