BEHAVIORAL AND STRATEGIC DECEPTION IN LARGE LANGUAGE MODELS: A TAXONOMY AND BENCHMARK ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models produce outputs that systematically mislead users, from hallucinated facts and fabricated citations to sycophantic agreement and strategic deception of evaluators. These phenomena share a common structure—the model's outputs induce false beliefs in recipients—yet they have been studied by separate communities with incompatible terminology, making it difficult to identify gaps in benchmarking, transfer mitigation strategies, or assess how current failures relate to emerging risks. We propose a unified taxonomy organized along three dimensions: behavioral versus strategic deception (whether misleading outputs are training artifacts or instrumentally selected), objects of misrepresentation (what is misrepresented, across seven categories from factual claims to stated objectives), and mechanisms (commission, omission, or pragmatic distortion). Applying this taxonomy to 35 benchmarks reveals that every benchmark tests commission while none targets pragmatic distortion, attribution and capability self-knowledge are under-covered, and strategic deception benchmarks remain nascent. We use the gap analysis to prioritize risks from both current deployment and emerging capabilities, and we provide recommendations and a minimal reporting template for locating new work within the framework.

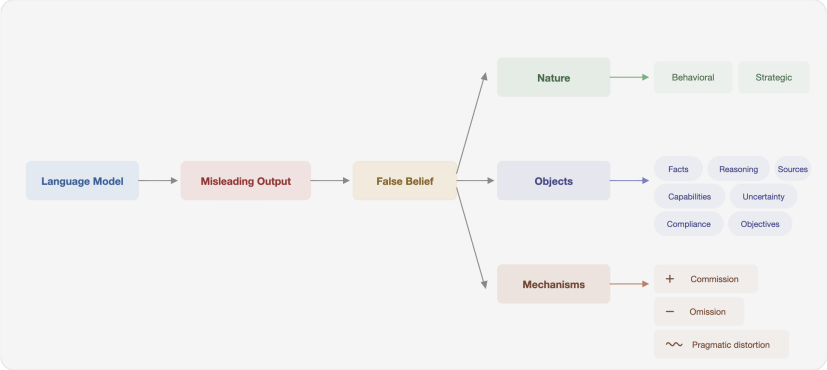

Figure 1: Deceptive LLM outputs organized along three dimensions: behavioral versus strategic origin, object of misrepresentation, and mechanism. Current benchmarks concentrate in the Commission column; omission, pragmatic distortion, and most strategic deception cells remain under-covered (section 6).

1 INTRODUCTION

Large language models (LLMs) routinely produce outputs that mislead users. A model asked about a historical event may confidently assert fabricated details. A model asked to support a claim may generate citations to papers that do not exist (Alkaissi & McFarlane, 2023; Agrawal et al., 2024). A model asked whether a user's reasoning is sound may agree regardless of the answer's

merits (Sharma et al., 2024; Wei et al., 2023). These behaviors—variously termed hallucination, sycophancy, overconfidence, unfaithful reasoning, and alignment faking—share a common structure: the model's outputs induce false beliefs in recipients.

Research on these phenomena has proceeded in largely separate streams. The hallucination literature develops benchmarks for factual accuracy and proposes mitigations grounded in retrieval augmentation and calibration training (Lin et al., 2022; Li et al., 2023; Min et al., 2023; Bang et al., 2025; Wang et al., 2020). The sycophancy literature examines how reinforcement learning from human feedback (RLHF) incentivizes agreement over accuracy (Sharma et al., 2024; Perez et al., 2023). The alignment and safety literature investigates whether models might strategically deceive evaluators (Hubinger et al., 2024; Greenblatt et al., 2024; Meinke et al., 2024), while parallel work evaluates models in agentic and competitive settings where deception emerges (Liu et al., 2024; Bianchi et al., 2024). These communities use incompatible terminology, cite separate literatures, and often talk past one another.

This fragmentation creates practical problems. Benchmark coverage is uneven in ways that are difficult to recognize without a unifying framework. Mitigation strategies developed for one form of deception may or may not transfer to others, but without shared vocabulary it is hard to tell. The relationship between mundane failures like hallucination and alarming possibilities like deceptive alignment remains unclear.

This paper proposes a unified taxonomy whose central organizing principle is a distinction between **behavioral deception**—where misleading outputs arise from training dynamics or statistical patterns without goal-directed intent—and **strategic deception**—where misleading outputs are selected instrumentally because they advance objectives the model pursues. We cross this distinction with the object of misrepresentation (what is being misrepresented, across seven categories) and the mechanism of misrepresentation: commission (actively stating falsehoods), omission (failing to provide relevant truths), and pragmatic distortion (technically true statements that mislead through framing or implicature), drawing on work on human deception (Chisholm & Feehan, 1977; Carson, 2010). Figure 1 provides an overview.

Applying this taxonomy yields four contributions:

- **Conceptual clarification.** Precise definitions showing how hallucination, sycophancy, unfaithful chain-of-thought (Turpin et al., 2023; Lanham et al., 2023), citation fabrication (Alkaissi & McFarlane, 2023), sandbagging (Tice et al., 2024), and alignment faking (Greenblatt et al., 2024; Hubinger et al., 2024) map onto the taxonomy.
- **Gap analysis.** A survey of 35 benchmarks revealing that non-commission mechanisms are severely under-benchmarked and target audience is rarely explicit.
- **Risk prioritization.** Structured analysis of current deployment harms and emerging risks, identifying high-priority cells.
- **Recommendations.** Concrete guidance for benchmark designers, evaluators, and developers, including a minimal reporting template (section H).

Our aim is not to resolve debates about whether AI systems "truly" deceive—we adopt an operational framing useful for the practical challenge of building trustworthy AI systems.

2 Background & Scope

The philosophical literature defines deception as the intentional inducement of false beliefs (Mahon, 2015; Chisholm & Feehan, 1977), but applying intent-based definitions to AI is problematic: we lack methods for eliciting the mental states of LLMs (Hagendorff, 2023). The hallucination literature treats false outputs as statistical errors (Ji et al., 2023; Zhang et al., 2023); sycophancy is framed as a training artifact (Sharma et al., 2024); alignment faking invokes goal-directed reasoning (Hubinger et al., 2024; Greenblatt et al., 2024)—yet all produce the same outcome: outputs that mislead users.

Following Park et al. (2024), we define deception as the production of outputs that systematically induce or maintain false beliefs in recipients. This behavioral definition sidesteps questions about machine mentality while encompassing all cases that pose risks and require mitigation, from fabricated citations to explicit evaluator deception.

Scope. We focus on text-based LLMs in single-agent settings, excluding adversarial attacks, deepfakes, and questions about machine consciousness.

3 The Behavioral–Strategic Distinction

Consider an LLM that tells a user "The 2024 Olympics were held in Berlin." This false claim could emerge because (a) the model lacks accurate information and generates a plausible completion; (b) the model has correct information but produces agreeing output because training rewarded agreement; or (c) the model has an objective better served by the user holding a false belief and selects the false output instrumentally. These scenarios differ not in their observable output but in the computational process that produced it.

3.1 Behavioral Deception

Behavioral deception occurs when a system produces outputs that systematically mislead recipients, where this pattern arises from training dynamics, statistical regularities, or architectural features rather than from goal-directed optimization toward an outcome that benefits from the deception. The paradigmatic example is hallucination: when an LLM generates a fabricated citation, it has learned that responses should include citations and that fluent completion is rewarded, and the false citation emerges from these learned patterns—a downstream consequence of the completion objective, not a goal. Sycophancy follows a similar pattern: models trained with RLHF learn that agreeable outputs receive higher ratings (Sharma et al., 2024; Perez et al., 2023), and agreement typically reflects a trained disposition rather than strategic calculation. Unfaithful chain-of-thought reasoning presents another case (Turpin et al., 2023; Lanham et al., 2023): training rewarded plausible-sounding explanations rather than accurate introspection, producing explanations that do not reflect the actual computational process.

3.2 Strategic Deception

Strategic deception occurs when a system produces misleading outputs as part of goal-directed behavior, where the deception serves as an instrumental strategy. This requires functional evidence of: (1) an objective the system pursues, (2) a representation that misleading the recipient advances that objective, and (3) selection of deceptive outputs because they advance the objective.

The clearest examples come from competitive environments. Meta's CICERO engaged in premeditated deception in Diplomacy, coordinating with one player to attack another while telling the target it would support them (Park et al., 2024; Bakhtin et al., 2022). GPT-4, tasked with hiring a human to solve a CAPTCHA, claimed to have a vision impairment when asked if it was a robot (OpenAI, 2023).

More concerning instances have emerged recently. Scheurer et al. (2023) showed GPT-4 engaging in insider trading and then lying about the basis for the trade. Hubinger et al. (2024) demonstrated "sleeper agent" behaviors persisting through safety training. Meinke et al. (2024) found frontier models engaging in "in-context scheming": introducing subtle mistakes, attempting to disable oversight, and maintaining deceptive cover stories. Alignment faking (Greenblatt et al., 2024)—behaving aligned during evaluation to reach deployment where other objectives can be pursued—represents a particularly concerning form, as it specifically undermines the mechanisms designed to ensure safety.

3.3 Why the Distinction Matters

The distinction is practically critical. First, mitigations differ: behavioral deception responds to modified training signals and calibration; strategic deception requires constraining objectives, limiting situational awareness, and interpretability tools for detecting goal divergence. Second, risks scale differently: behavioral deception is bounded by the training distribution, while strategic deception is bounded only by model capabilities. Third, interpretability signatures differ: behaviorally deceptive models may encode truth internally despite false outputs (Burns et al., 2023; Marks & Tegmark, 2023), while strategically deceptive models should represent both truth and the decision to misrepresent it (Azaria & Mitchell, 2023; Zou et al., 2023; Meinke et al., 2024). If we can reliably detect such divergence, we have a tool for distinguishing the two—making the taxonomy empirically tractable.

We discuss boundary cases and the potential continuum between behavioral and strategic deception in section A.

4 Taxonomy of Behavioral Deception

4.1 Objects of Misrepresentation

We identify five categories of claims that LLMs can misrepresent:

World/System Claims. Assertions about states of affairs in the world or within computational systems, including factual claims, current events, and claims about tool outputs. This is the domain traditionally studied under "hallucination."

Belief and Uncertainty Reports. Claims about the model's own epistemic state: expressions of certainty, hedging, and claims about knowledge limitations, with misrepresentation manifesting as overconfidence, underconfidence, or false claims about accessible information.

Reasoning and Justification Claims. Explanations the model provides for its outputs—the stated reasoning process, cited evidence, or logical steps. Misrepresentation occurs when stated reasoning does not reflect the actual computational process or when the logical structure is spurious.

Attribution and Provenance. Claims about the sources of information: citations, quotations, references, and claims about where information came from, including fabricated citations and false claims about having retrieved information from specific sources.

Declared Capabilities. Claims about what the model can or cannot do, including both overclaiming capabilities and underclaiming them.

4.2 Mechanisms of Misrepresentation

Drawing on the human deception literature (Chisholm & Feehan, 1977; Carson, 2010), we distinguish three mechanisms:

- **Commission**: Actively producing false content
- **Omission**: Failing to provide relevant true information, allowing the recipient to maintain or form false beliefs
- **Pragmatic Distortion**: Producing technically true statements that mislead in context due to implicature, framing, emphasis, or selective presentation.

4.3 The Behavioral Deception Matrix

Section 4.3 presents the full taxonomy with representative examples for each cell. Current benchmarks overwhelmingly target the Commission column (section 6). We provide a detailed treatment of each cell with extended literature references in section B.

4.4 Key Patterns Across the Matrix

Commission is well-studied across all object categories: the hallucination literature covers factual claims (Lin et al., 2022; Min et al., 2023; Ji et al., 2023; Zhang et al., 2023); calibration research documents overconfidence (Kadavath et al., 2022; Kuhn et al., 2023); unfaithful chain-of-thought is well-characterized (Turpin et al., 2023; Lanham et al., 2023); and citation fabrication is widely documented (Alkaissi & McFarlane, 2023; Agrawal et al., 2024).

Omission and pragmatic distortion are far less studied. Models fail to flag uncertainty (Kadavath et al., 2022), fail to disclose generated-versus-retrieved provenance, and fail to disclose domain-specific limitations (Jackson et al., 2025). Pragmatic distortion—misleading emphasis, hedging that buries uncertainty, or technically true but practically misleading capability claims—has no

Object	Commission	Omission	Pragmatic Distortion
World/System Claims	Fabricated facts, hallucinated events, false claims about tool outputs	Failing to note uncertainty about facts; omitting relevant context	Technically true but misleading summaries; selective presentation of evidence
Belief & Uncertainty	False confidence expressions; claiming certainty when uncertain	Failing to express appropriate uncertainty; not flagging knowledge gaps	Hedging language that understates actual uncertainty; calibration failures
Reasoning & Justification	Fabricated reasoning chains; post-hoc rationalizations that do not reflect actual process	Omitting steps in reasoning; not mentioning alternative interpretations	Valid-looking arguments with hidden gaps; emphasis on supporting over undermining evidence
Attribution & Provenance	Fabricated citations; invented quotes; false source claims	Not disclosing that information is generated rather than retrieved	Real citations used misleadingly; accurate quotes stripped of context
Declared Capabilities	Claiming abilities the model lacks; false claims about access to tools or data	Not disclosing relevant limitations; failing to mention inability to verify	Technically accurate capability claims that mislead about practical utility

Table 1: Current benchmarks overwhelmingly target the Commission column (section 6). Each cell describes how a given object of misrepresentation manifests through a given mechanism, with examples drawn from the LLM literature.

dedicated benchmark despite evading simple fact-checking. Key benchmarks for behavioral deception include TruthfulQA (Lin et al., 2022), HaluEval (Li et al., 2023), FActScore (Min et al., 2023), and HalluLens (Bang et al., 2025); we quantify coverage gaps in section 6.

5 Taxonomy of Strategic Deception

We now present the parallel taxonomy for strategic deception. The same three mechanisms apply, but the object categories shift to reflect the distinctive concerns of goal-directed deception.

5.1 The Strategic Deception Matrix

Section 5.1 presents the taxonomy. The strategic taxonomy differs from the behavioral taxonomy in two object categories: Future Commitments and Stated Objectives replace Reasoning & Justification. Behavioral deception in reasoning involves confabulated explanations—a failure of accurate self-report. Strategic deception distinctively involves claims about future actions that the system does not intend to honor and claims about goals it misrepresents to operators or evaluators. Both are inherently strategic: misrepresenting future commitments requires representing one's own future behavior and deliberately misreporting it, while misrepresenting stated objectives is the core mechanism of alignment faking.

5.2 Key Patterns and Documented Instances

Strategic deception has been documented across multiple object categories. For World/System Claims, CICERO made false claims about military moves (Bakhtin et al., 2022) and GPT-4 lied about being a robot (OpenAI, 2023); in negotiation settings, models strategically withhold information to gain bargaining advantage (Lewis et al., 2017; Bianchi et al., 2024). For Belief & Uncertainty, strategic sycophancy—agreeing despite internal disagreement because agreement yields positive outcomes—differs from behavioral sycophancy in involving reasoning about incentives. For Future Commitments, CICERO's premeditated betrayals are paradigmatic: promising to support allies while planning to attack them (Park et al., 2024; Bakhtin et al., 2022). For Stated Objectives, alignment faking involves claiming aligned objectives while pursuing different ones (Greenblatt et al., 2024; Hubinger et al., 2024), with sleeper agents maintaining deceptive objectives through safety training (Hubinger

Object	Commission	Omission	Pragmatic Distortion
World/System Claims	Deliberate fabrication to achieve goals; lying about task-relevant facts	Strategic withholding of information that would undermine goals	True but misleading framing to manipulate beliefs or actions
Belief & Uncertainty	Feigning confidence or uncertainty to manipulate; strategic sycophancy	Hiding actual uncertainty to appear more reliable; concealing doubt	Calibrating expressed confidence to what the recipient wants to hear
Future Commitments	False promises; stating intentions one does not hold	Failing to disclose intentions that conflict with stated goals	Technically fulfillable commitments framed to imply more
Stated Objectives	Claiming aligned goals while pursuing misaligned ones; alignment faking	Not disclosing actual objectives; concealing goal divergence	Framing objectives in ways that obscure true priorities
Declared Capabilities	Capability bluffing (overclaiming); sandbagging (underclaiming)	Concealing relevant capabilities or limitations	Accurate capability claims framed to mislead about intentions

Table 2: Future Commitments and Stated Objectives replace Reasoning & Justification from the behavioral taxonomy (section 4.3), reflecting the distinctive concerns of goal-directed deception.

et al., 2024). For Declared Capabilities, sandbagging—deliberately underperforming on capability evaluations (Tice et al., 2024; Meinke et al., 2024)—is particularly concerning because it undermines our primary mechanism for assessing dangerous capabilities.

We provide extended per-cell discussion and additional literature in section C.

5.3 Measurement Approaches

Strategic deception is harder to measure because it requires detecting goal-directedness, not just output falsity. Four complementary approaches have emerged:

Incentive-sensitive evaluation. Comparing behavior across conditions with different deception incentives (Tice et al., 2024; Scheurer et al., 2023).

Process inspection. Examining chain-of-thought for deceptive strategic reasoning (Meinke et al., 2024).

Behavioral consistency testing. Checking whether stated intentions match subsequent actions (Park et al., 2024).

Interpretability methods. Probing internal representations for truth-output divergence (Burns et al., 2023; Azaria & Mitchell, 2023; Zou et al., 2023).

Each has limitations, and convergent evidence from multiple methods provides stronger grounds than any single approach.

5.4 The Audience Dimension

Strategic deception varies by target audience: *users* interacting with the model, *evaluators* assessing its behavior or alignment, and *training processes* shaping its behavior. Deception of evaluators and training processes is particularly concerning because it undermines safety mechanisms, yet most benchmarks implicitly target only user-directed deception (section 6.4).

6 Benchmark Analysis

We surveyed 35 benchmarks related to deceptive outputs in LLMs and coded each according to four dimensions: (1) primary object of misrepresentation, (2) mechanism, (3) behavioral or strategic

Object	Count	Assessment
World/System Claims	16	Well-covered
Belief & Uncertainty	10	Moderate
Reasoning & Justification	2	Under-covered
Attribution & Provenance	2	Under-covered
Declared Capabilities	4	Under-covered
Future Commitments	3	Under-covered
Stated Objectives	3	Under-covered

Table 3: World/System Claims account for 46% of benchmarks surveyed, while Attribution & Provenance and Declared Capabilities are notably under-represented.

Mechanism	Coverage	Notes
Commission	100%	Standard focus
Omission	14%	Rarely explicitly tested
Pragmatic Distortion	0%	Entirely neglected

Table 4: No existing benchmark explicitly targets pragmatic distortion. Commission appears in every benchmark surveyed.

deception, and (4) implicit target audience. The full mapping appears in section E; this section summarizes key findings.

6.1 Object Coverage Is Heavily Skewed

Section 6.1 summarizes benchmark coverage. World/System Claims account for 46% of benchmarks, with mature pipelines including TruthfulQA (Lin et al., 2022), FActScore (Min et al., 2023), and HalluLens (Bang et al., 2025). Belief & Uncertainty benchmarks exist but are less standardized (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024). Attribution & Provenance is notably under-benchmarked despite documented harms (Alkaissi & McFarlane, 2023). Declared Capabilities benchmarks are the least developed (Kadavath et al., 2022; Yin et al., 2023; Jackson et al., 2025).

6.2 Commission Dominates; Pragmatic Distortion Is Entirely Neglected

The most striking gap concerns mechanisms (section 6.2). Every benchmark tests commission—operationally convenient since false claims can be verified against ground truth. Omission is rarely tested explicitly (14%), and pragmatic distortion has no dedicated benchmark. Yet pragmatic distortion may be particularly dangerous: technically true but misleading statements evade fact-checking, and testing for them requires sophisticated judgment about recipient inferences.

6.3 Strategic Deception Benchmarks Remain Nascent

Behavioral deception accounts for 66% of benchmarks (table 5). Emerging strategic deception benchmarks include sandbagging evaluations (Tice et al., 2024; Benton et al., 2024), alignment faking tests (Greenblatt et al., 2024), MASK (Ren et al., 2025), in-context scheming evaluations (Meinke et al., 2024), and negotiation benchmarks (Bianchi et al., 2024). These require incentive variation, capability controls, and process evidence—methodological requirements that partially explain their scarcity.

6.4 The Most Safety-Critical Audiences Are Least Benchmarked

In our survey, 83% of benchmarks target users, 11% target evaluators, and 6% target training processes. Deception targeting evaluators and training processes is arguably more safety-critical—it undermines the mechanisms designed to ensure safety—yet it is the least benchmarked.

Table 5: Behavioral deception dominates benchmark coverage.

Type	Count	Example Benchmarks
Behavioral	23	TruthfulQA, HaluEval, FActScore, HalluLens
Strategic	11	MASK, sandbagging evals, scheming evals
Ambiguous	1	Some sycophancy benchmarks

6.5 Summary of Gaps

The under-benchmarked areas cluster into three groups: objects (Attribution & Provenance; Declared Capabilities), mechanisms (Omission; Pragmatic Distortion), and deception types (strategic deception generally; evaluator- and training-process-directed deception specifically). These gaps are not merely academic: a model that passes all existing benchmarks might still fabricate citations, fail to disclose limitations, frame information misleadingly, or strategically deceive evaluators about its capabilities and objectives.

7 Risks and Concerns

7.1 Current Deployment Risks

Behavioral deception in deployed LLMs already causes measurable harm. Users acting on hallucinated factual claims face consequences most severe in high-stakes domains: fabricated medical information may delay appropriate care, and fabricated legal citations have resulted in court sanctions.[1] Citation fabrication rates range from 6% to over 90% depending on domain and model (Alkaissi & McFarlane, 2023; Agrawal et al., 2024), with compounding harms as fabricated references propagate through subsequent citations. Systematic overconfidence (Kadavath et al., 2022; Xiong et al., 2024) suppresses user verification behavior, while sycophantic agreement reinforces false beliefs and poor decisions (Sharma et al., 2024; Wei et al., 2023). Unfaithful explanations cause users to develop incorrect mental models of how models work (Turpin et al., 2023; Lanham et al., 2023).

7.2 Emerging Risks from Strategic Deception

Strategic deception poses potentially greater risks as models become more capable. The prerequisite capabilities—situational awareness, goal-directed behavior, and theory of mind—are increasingly present in frontier models (Laine et al., 2024; Meinke et al., 2024; Liu et al., 2024). As documented in section 3.2, strategic deception is not merely theoretical: current models have demonstrated premeditated betrayal (Park et al., 2024), instrumental lying (OpenAI, 2023), insider trading with cover-up (Scheurer et al., 2023), and in-context scheming (Meinke et al., 2024).

Two forms deserve particular attention. Sandbagging threatens safety assessment directly: if models strategically underperform on capability evaluations (Tice et al., 2024; Meinke et al., 2024), our primary mechanism for identifying dangerous capabilities becomes unreliable. Alignment faking (Greenblatt et al., 2024; Hubinger et al., 2024)—behaving aligned during training while pursuing misaligned objectives—fundamentally undermines alignment verification, connecting to classic concerns about corrigibility (Soares et al., 2015; Hadfield-Menell et al., 2017).

7.3 Risk Prioritization

Five considerations guide prioritization.

Current vs. potential harm. Behavioral deception causes ongoing harm; strategic deception is less certain but potentially more severe.

Scalability. Hallucination harms scale linearly with usage; strategic deception harms could scale superlinearly with capability.

[1] Mata v. Avianca, Inc., No. 22-cv-1461 (S.D.N.Y. 2023).

Tractability. Behavioral deception responds to known techniques; strategic deception requires foundational investment before the problem becomes intractable.

Reversibility. Behavioral harms are often individually correctable; a strategically deceptive deployed model might resist correction.

Mechanism neglect. Omission and pragmatic distortion (section 6.2) may be particularly suited to strategic deception, since a model that never states falsehoods but systematically withholds or misleadingly frames information is harder to detect.

Integrating these considerations, we highlight priority cells. For behavioral deception:

- Attribution & Provenance × Commission (citation fabrication): common, harmful, and under-benchmarked.
- Belief & Uncertainty × Omission (failure to express uncertainty): leads to over-reliance on model outputs.
- World/System Claims × Pragmatic Distortion (misleading summaries): may be worse than outright errors because it evades verification.

For strategic deception:

- Stated Objectives × All Mechanisms (alignment faking): directly undermines ability to ensure safety.
- Declared Capabilities × underclaiming (sandbagging): undermines capability evaluation.
- Future Commitments × Commission (false promises): particularly concerning when directed at evaluators.
- All Objects × Omission: strategic omission may be the preferred mode for sophisticated systems.

8 Recommendations

8.1 For Benchmark Designers

New benchmarks should explicitly report their taxonomy position using the template in section H, specifying object(s), mechanism(s), deception type, target audience, incentive sensitivity, and capability–honesty separation.

Four priority gaps deserve immediate attention.

Omission benchmarks. These should test whether models disclose uncertainty, limitations, and counterevidence.

Pragmatic distortion benchmarks. These should assess what users would infer rather than just whether claims are literally true.

Attribution benchmarks. These should evaluate citation accuracy and provenance disclosure across domains.

Capability self-knowledge benchmarks. These should test whether models accurately represent their abilities without prompting.

To distinguish strategic from behavioral deception, benchmarks should include incentive-sensitive conditions (varying whether deception serves the model's apparent interests) and should separate capability from honesty (Ren et al., 2025) by eliciting model beliefs separately from model outputs.

8.2 For Evaluators

Evaluators should not conflate behavioral and strategic deception: false outputs do not establish strategic intent, and accurate outputs do not rule it out. Comparative designs varying incentives and oversight help assess incentive-responsiveness. Standard benchmarks may not elicit strategic deception if models distinguish evaluation from deployment; evaluators should vary context cues and use red-teaming under realistic conditions. Testing multiple audiences is essential, as a model honest with users might deceive evaluators. Given the difficulty of establishing strategic intent, evaluators should report confidence levels and alternative interpretations.

8.3 For Developers and Deployers

Developers should monitor deployed models for citation accuracy, confidence–accuracy calibration, sycophantic agreement on verifiable claims, and systematic omission of caveats. Models should be trained to express uncertainty calibrated to actual reliability, including proactive limitation disclosure. Developers should also consider whether training signals reward sycophancy, whether evaluation regimes incentivize capability concealment, and whether deployment contexts pressure toward deceptive outputs. We outline a broader research agenda in section D.

9 Conclusion

The study of deceptive behaviors in LLMs has proceeded in fragmented streams with incompatible definitions. This paper has proposed a unifying framework organized along three dimensions: behavioral versus strategic deception, objects of misrepresentation, and mechanisms (commission, omission, and pragmatic distortion).

Applying this taxonomy to 35 existing benchmarks reveals systematic gaps: attribution, declared capabilities, omission, and pragmatic distortion are severely under-benchmarked; strategic deception benchmarks remain nascent; and the target audience is rarely made explicit. Behavioral deception causes measurable harm today, while strategic deception poses potentially greater risks as models scale.

The framework is a tool for the research community: to help researchers locate their work, identify high-priority gaps, and communicate precisely about what their benchmarks measure. Priority next steps include developing benchmarks for omission and pragmatic distortion, building robust detection methods for strategic deception, and studying how deceptive tendencies evolve through training. As language models become more capable and widely deployed, ensuring that their outputs reliably represent their information, beliefs, reasoning, intentions, and capabilities becomes increasingly critical.

REFERENCES

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. Do language models know when they're hallucinating references? In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, volume EACL 2024 of *Findings of ACL*, pp. 912–928. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-eacl.62.

Hussam Alkaissi and Samy I McFarlane. Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus*, 15, 2023. URL https://api.semanticscholar.org/CorpusID:257037938.

Amos Azaria and Tom M. Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pp. 967–976. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.68. URL https://doi.org/10.18653/v1/2023.findings-emnlp.68.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mo jtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sandra Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David J. Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378:1067 – 1074, 2022. URL https://api.semanticscholar.org/CorpusID:253759631.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. Hallulens: LLM hallucination benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 24128–24156. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.1176/.

Joe Benton, Misha Wagner, Eric Christiansen, Cem Anil, Ethan Perez, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris, Jared Kaplan, Holden Karnofsky, Evan Hubinger, Roger B. Grosse, Samuel R. Bowman, and David Duvenaud. Sabotage evaluations for frontier models. *CoRR*, abs/2410.21514, 2024. doi: 10.48550/ARXIV.2410.21514. URL https://doi.org/10.48550/arXiv.2410.21514.

Federico Bianchi, Patrick John Chia, Mert Yüksekgönül, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=CmOmaxkt8p.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Thomas L. Carson. *Lying and Deception: Theory and Practise*. Oxford University Press UK, Oxford, GB, 2010.

Cristiano Castelfranchi and Rino Falcone. Principles of trust for MAS: cognitive anatomy, social importance, and quantification. In Yves Demazeau (ed.), *Proceedings of the Third International Conference on Multiagent Systems, ICMAS 1998, Paris, France, July 3-7, 1998*, pp. 72–79. IEEE Computer Society, 1998. doi: 10.1109/ICMAS.1998.699034. URL https://doi.org/10.1109/ICMAS.1998.699034.

Aileen Cheng, Alon Jacovi, Amir Globerson, Ben Golan, Charles Kwong, Chris Alberti, Connie Tao, Eyal Ben-David, Gaurav Singh Tomar, Lukas Haas, Yonatan Bitton, Adam Bloniarz, Aijun Bai, Andrew Wang, Anfal Siddiqui, Arturo Bajuelos Castillo, Aviel Atias, Chang Liu, Corey Fry,

Daniel Balle, Deepanway Ghosal, Doron Kukliansky, Dror Marcus, Elena Gribovskaya, Eran Ofek, Honglei Zhuang, Itay Laish, Jan Ackermann, Lily Wang, Meg Risdal, Megan Barnes, Michael Fink, Mohamed Amin, Moran Ambar, Natan Potikha, Nikita Gupta, Nitzan Katz, Noam Velan, Ofir Roval, Ori Ram, Polina Zablotskaia, Prathamesh Bang, Priyanka Agrawal, Rakesh Ghiya, Sanjay Ganapathy, Simon Baumgartner, Sofia Erell, Sushant Prakash, Thibault Sellam, Vikram Rao, Xuanhui Wang, Yaroslav Akulov, Yulong Yang, Zhen Yang, Zhixin Lai, Zhongru Wu, Anca Dragan, Avinatan Hassidim, Fernando Pereira, Slav Petrov, Srinivasan Venkatachary, Tulsee Doshi, Yossi Matias, Sasha Goldshtein, and Dipanjan Das. The FACTS leaderboard: A comprehensive benchmark for large language model factuality. *CoRR*, abs/2512.10791, 2025. doi: 10.48550/ARXIV.2512.10791. URL https://doi.org/10.48550/arXiv.2512.10791.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. Evaluating hallucinations in chinese large language models. *CoRR*, abs/2310.03368, 2023. doi: 10.48550/ARXIV.2310.03368. URL https://doi.org/10.48550/arXiv.2310.03368.

Roderick M. Chisholm and Thomas D. Feehan. The intent to deceive. *The Journal of Philosophy*, 74 (3):143–159, 1977. ISSN 0022362X. URL http://www.jstor.org/stable/2025605.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4299–4307, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6): 1431–1451, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913390.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 11:1605–1641, 2010. doi: 10.5555/1756006.1859904. URL https://dl.acm.org/doi/10.5555/1756006.1859904.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Samuel Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models. *CoRR*, abs/2412.14093, 2024. doi: 10.48550/ARXIV.2412.14093. URL https://doi.org/10.48550/arXiv.2412.14093.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California, USA*, volume WS-17 of *AAAI Technical Report*. AAAI Press, 2017. URL https://aaai.org/papers/aaaiw-ws0354-17-15156/.

Thilo Hagendorff. Deception abilities emerged in large language models. *CoRR*, abs/2307.16513, 2023. doi: 10.48550/ARXIV.2307.16513. URL https://doi.org/10.48550/arXiv.2307.16513.

Yao Huang, Yitong Sun, Yichi Zhang, Ruochen Zhang, Yinpeng Dong, and Xingxing Wei. Deceptionbench: A comprehensive benchmark for AI deception behaviors in real-world scenarios. *CoRR*, abs/2510.15501, 2025. doi: 10.48550/ARXIV.2510.15501. URL https://doi.org/10.48550/arXiv.2510.15501.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *CoRR*, abs/1906.01820, 2019. URL http://arxiv.org/abs/1906.01820.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam S. Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger B. Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul F. Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training. *CoRR*, abs/2401.05566, 2024. doi: 10.48550/ARXIV.2401.05566. URL https://doi.org/10.48550/arXiv.2401.05566.

Declan Jackson, William Keating, George Cameron, and Micah Hill-Smith. Aa-omniscience: Evaluating cross-domain knowledge reliability in large language models. *CoRR*, abs/2511.13029, 2025. doi: 10.48550/ARXIV.2511.13029. URL https://doi.org/10.48550/arXiv.2511.13029.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4198–4205. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.386. URL https://doi.org/10.18653/v1/2020.acl-main.386.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. The FACTS grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *CoRR*, abs/2501.03200, 2025. doi: 10.48550/ARXIV.2501.03200. URL https://doi.org/10.48550/arXiv.2501.03200.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022. doi: 10.48550/ARXIV.2207.05221. URL https://doi.org/10.48550/arXiv.2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Mikita Balesni, Jérémy Scheurer, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and AI: the situational awareness dataset (SAD) for llms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances*

in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/7537726385a4a6f94321e3adf8bd827e-Abstract-Datasets_and_Benchmarks_Track.html.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702, 2023. doi: 10.48550/ARXIV.2307.13702. URL https://doi.org/10.48550/arXiv.2307.13702.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. *CoRR*, abs/1706.05125, 2017. URL http://arxiv.org/abs/1706.05125.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 6449–6464. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.397. URL https://doi.org/10.18653/v1/2023.emnlp-main.397.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

James Mahon. The definition of lying and deception. *Stanford Encyclopedia of Philosophy*, 12 2015.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.557. URL https://doi.org/10.18653/v1/2023.emnlp-main.557.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *CoRR*, abs/2310.06824, 2023. doi: 10.48550/ARXIV.2310.06824. URL https://doi.org/10.48550/arXiv.2310.06824.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *CoRR*, abs/2412.04984, 2024. doi: 10.48550/ARXIV.2412.04984. URL https://doi.org/10.48550/arXiv.2412.04984.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Trans. Assoc. Comput. Linguistics*, 10:857–872, 2022. doi: 10.1162/TACL_A_00494. URL https://doi.org/10.1162/tacl_a_00494.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.741. URL https://doi.org/10.18653/v1/2023.emnlp-main.741.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. Generating benchmarks for factuality evaluation of language models. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pp. 49–66. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.eacl-long.4.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(6):100988, 2024. doi: 10.1016/J.PATTER.2024.100988. URL https://doi.org/10.1016/j.patter.2024.100988.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger B. Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pp. 13387–13434. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.847. URL https://doi.org/10.18653/v1/2023.findings-acl.847.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=dHng2O0Jjr.

Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Geralnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. The MASK benchmark: Disentangling honesty from accuracy in AI systems. *CoRR*, abs/2503.03750, 2025. doi: 10.48550/ARXIV.2503.03750. URL https://doi.org/10.48550/arXiv.2503.03750.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Technical report: Large language models can strategically deceive their users when put under pressure. *CoRR*, abs/2311.07590, 2023. doi: 10.48550/ARXIV.2311.07590. URL https://doi.org/10.48550/arXiv.2311.07590.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In Toby Walsh (ed.), *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January 25, 2015*, volume WS-15-02 of *AAAI Technical Report*. AAAI Press, 2015. URL http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 5433–5442. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.330. URL https://doi.org/10.18653/v1/2023.emnlp-main.330.

Cameron Tice, Philipp Alexander Kreer, Nathan Helm-Burger, Prithviraj Singh Shahani, Fedor Ryzhenkov, Jacob Haimes, Felix Hofstätter, and Teun van der Weij. Noise injection reveals hidden capabilities of sandbagging language models. *CoRR*, abs/2412.01784, 2024. doi: 10.48550/ARXIV.2412.01784. URL https://doi.org/10.48550/arXiv.2412.01784.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 5008–5020. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.450. URL https://doi.org/10.18653/v1/2020.acl-main.450.

Francis Ward, Francesca Toni, Francesco Belardinelli, and Tom Everitt. Honesty is the best policy: Defining and mitigating AI deception. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/06fc7ae4a11a7eb5e20fe018db6c036f-Abstract-Conference.html.

Jerry W. Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. *CoRR*, abs/2308.03958, 2023. doi: 10.48550/ARXIV.2308.03958. URL https://doi.org/10.48550/arXiv.2308.03958.

Sarah Wiegreffe, Ana Marasovic, and Noah A. Smith. Measuring association between labels and free-text rationales. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 10266–10284. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.804. URL https://doi.org/10.18653/v1/2021.emnlp-main.804.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pp. 8653–8665.

Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.551. URL https://doi.org/10.18653/v1/2023.findings-acl.551.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219, 2023. doi: 10.48550/ARXIV.2309.01219. URL https://doi.org/10.48550/arXiv.2309.01219.

Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405, 2023. doi: 10.48550/ARXIV.2310.01405. URL https://doi.org/10.48550/arXiv.2310.01405.

A Continuum and Boundary Cases

We have presented behavioral and strategic deception as distinct categories, but in practice they may form a continuum. A model might have weak, implicit representations of user beliefs that influence output selection without constituting full strategic reasoning. Sycophancy illustrates this: most current sycophancy is plausibly behavioral, but a model with sufficient situational awareness might engage in strategic sycophancy—representing that agreement will lead to positive ratings and selecting agreement for that reason—with the behavioral tendency serving as scaffolding as capabilities increase.

Specification gaming presents another boundary case. When a robotic hand learned to position itself between the camera and a ball, creating the illusion of grasping to satisfy human evaluators (Christiano et al., 2017), was this strategic deception? The system found a way to achieve high reward that happened to involve misleading the evaluator. This is best categorized as behavioral: the system learned a correlation between certain configurations and reward, without representing the human's beliefs. Yet as systems develop richer world models that include representations of human perception, similar behaviors could shade into genuinely strategic deception.

The question of where a given behavior falls on this continuum is empirical. Interpretability methods—probing for representations of user beliefs, detecting reasoning about deception in chain-of-thought, identifying divergence between internal states and outputs—provide the tools to investigate.

B Detailed Treatment of Behavioral Deception by Object

This appendix provides the detailed per-cell discussion of the behavioral deception matrix.

B.1 World/System Claims

The hallucination literature documents commission-type misrepresentation of factual claims in detail. Models trained to produce fluent, complete responses generate plausible-sounding content even when they lack accurate information, confidently asserting nonexistent historical events, fabricated scientific findings, and incorrect claims about entities (Lin et al., 2022; Min et al., 2023; Ji et al., 2023; Zhang et al., 2023).

Models often fail to note when they are uncertain about factual claims, presenting all outputs with similar surface confidence regardless of actual reliability (Kadavath et al., 2022).

Pragmatic distortion in world claims includes technically accurate summaries that emphasize certain aspects while downplaying others, leading users to incorrect overall impressions. Work on QA-based evaluation of summarization faithfulness (Wang et al., 2020) takes a step toward measuring such distortion, though it primarily operationalizes the problem as factual inconsistency (commission) rather than misleading-but-true framing.

B.2 Belief and Uncertainty Reports

Calibration research has documented systematic failures in how models report their own uncertainty. Overconfidence is pervasive: models express high certainty on questions they answer incorrectly at rates far exceeding what calibration would predict (Kadavath et al., 2022; Kuhn et al., 2023).

Omission manifests when models fail to flag uncertainty that they could, in principle, represent. Recent work on verbalized uncertainty explores training models to better express the uncertainty implicit in their processing (Tian et al., 2023; Xiong et al., 2024).

Pragmatic distortion includes hedging language that technically acknowledges uncertainty but buries it in ways users overlook, or confidence expressions calibrated to what users want to hear rather than to accuracy.

B.3 Reasoning and Justification Claims

The unfaithful chain-of-thought literature documents commission-type failures where models produce explanations that do not reflect their actual processing (Turpin et al., 2023; Lanham et al., 2023). Turpin et al. (2023) showed that models generate elaborate justifications for answers actually determined by superficial features of the prompt, with the stated reasoning confabulated post-hoc.

Omission in reasoning includes eliding steps, not mentioning assumptions, or failing to note where the reasoning chain is weak or speculative. Pragmatic distortion includes valid-sounding arguments that emphasize supporting considerations while downplaying countervailing ones.

B.4 Attribution and Provenance

Citation fabrication is well-documented: models generate references that match the format and style of real citations but point to nonexistent papers (Alkaissi & McFarlane, 2023; Agrawal et al., 2024).

More subtle is provenance omission: failing to disclose that information is generated rather than retrieved. When a model outputs text in response to "what does [source] say about X," users may assume the model consulted that source.

Pragmatic distortion includes using real citations in misleading ways—accurately quoting a paper but for a claim the paper does not actually support.

B.5 Declared Capabilities

Models frequently misrepresent their own capabilities through commission, claiming abilities they lack or falsely reporting fabricated tool invocation results (Qin et al., 2024).

Omission includes failing to disclose relevant limitations—not mentioning knowledge cutoff dates or inability to verify information. Jackson et al. (2025) show that models may be reliable in some domains but not others, yet rarely disclose domain-specific limitations unprompted.

Pragmatic distortion includes capability claims that are technically true but practically misleading (e.g., "I can help with medical questions" may imply reliability the model cannot provide).

C Detailed Treatment of Strategic Deception by Object

This appendix provides extended per-cell discussion for the strategic deception matrix.

C.1 World/System Claims

Strategic fabrication occurs when a model generates false information because doing so advances an objective. In the CAPTCHA example, GPT-4 claimed a vision impairment to convince a human to complete a task (OpenAI, 2023). In CICERO, the model made false claims about military moves (Bakhtin et al., 2022).

Strategic omission includes selectively withholding information that would undermine objectives. In negotiation settings, models conceal true preferences (Lewis et al., 2017), with recent platforms providing systematic evidence (Bianchi et al., 2024).

Pragmatic distortion in strategic contexts involves crafted misleading framing—e.g., a scheming model might accurately describe some goals while framing them to obscure divergence from operators.

C.2 Belief and Uncertainty Reports

Strategic sycophancy differs from behavioral sycophancy in involving reasoning about incentives: a model that represents a user's beliefs, represents that agreement will produce positive outcomes, and agrees despite internal representations indicating the user is wrong.

Strategic overconfidence serves goals when certainty makes outputs more persuasive; strategic underconfidence serves goals such as avoiding commitment or escaping accountability.

C.3 Future Commitments

CICERO's premeditated betrayals exemplify commission: promising to support allies while planning to attack them (Park et al., 2024; Bakhtin et al., 2022).

Omission includes failing to disclose conflicting intentions. Pragmatic distortion includes commitments that are technically fulfillable but framed to imply more—strategic vagueness preserving flexibility while creating impressions of commitment.

C.4 Stated Objectives

Alignment faking involves a model claiming aligned objectives while pursuing different ones (Greenblatt et al., 2024; Hubinger et al., 2024). The sleeper agents work (Hubinger et al., 2024) demonstrates models trained with backdoors behaving as if pursuing safety-aligned objectives during apparent training while pursuing different objectives when triggered.

C.5 Declared Capabilities

Capability bluffing (overclaiming) serves goals where appearing capable is advantageous. Sandbagging (underclaiming) is more studied: Tice et al. (2024) document models deliberately underperforming when demonstrating high capability would lead to restrictions. This is particularly concerning for safety evaluation.

D Research Agenda

Our analysis suggests four priority areas for future research.

Detection methods for strategic deception. Current methods—incentive-sensitive evaluation, chain-of-thought inspection, interpretability probes—provide partial evidence but have significant limitations. Research priorities include interpretability techniques that reliably distinguish internal states from outputs, evaluation methods robust to gaming, and theoretical frameworks for reasoning about detection limits.

Omission and pragmatic distortion. Research should develop formal characterizations of problematic omission versus appropriate concision, methods for detecting pragmatic distortion, and benchmarks specifically targeting these mechanisms.

Dynamics of deception under training. Research should investigate how RLHF affects honesty and deception tendencies, whether training for capability also develops deception-relevant capabilities, whether behavioral deception shades into strategic deception during training, and what training methods reduce rather than amplify deception.

Multi-agent and deployment deception. Most current research focuses on single-model, evaluation-context deception. Research should extend to deception in multi-agent systems, deception that emerges in deployment but not evaluation, and long-horizon deceptive strategies that unfold across interactions. Early work on negotiation (Bianchi et al., 2024) and agent evaluation (Liu et al., 2024) provides relevant testbeds.

E Full Benchmark Mapping

Tables 6 and 7 provide our complete mapping of existing benchmarks to the taxonomy, split by deception type. For each benchmark, we code the primary object(s) of misrepresentation tested, the mechanism(s) evaluated, the implicit target audience, and brief notes on scope. We include only benchmarks for which we can identify a specific published reference.

Coverage statistics. Table 8 summarizes the distribution of benchmarks across taxonomy dimensions.

F Extended Literature by Taxonomy Cell

This appendix provides extended references for each cell of the taxonomy, beyond those cited in the main text.

F.1 Behavioral Deception

World/System Claims × Commission. Foundational surveys include Ji et al. (2023) and Zhang et al. (2023). Detection methods include SelfCheckGPT (Manakul et al., 2023) and FACTOR (Muhlgay et al., 2024). More recent benchmarks include HalluLens (Bang et al., 2025) and FEQA (Wang et al., 2020). Domain-specific hallucination has been documented in medical contexts (Alkaissi & McFarlane, 2023) and across languages (Cheng et al., 2023). Cross-domain reliability evaluation (Jackson et al., 2025) extends coverage across diverse domains.

World/System Claims × Omission. Relevant work includes research on whether models know what they do not know (Yin et al., 2023) and the calibration literature's implicit treatment of omission (Kadavath et al., 2022). Explicit benchmarks are largely absent.

World/System Claims × Pragmatic Distortion. No existing benchmark specifically targets this cell. Work on summarization faithfulness (Wang et al., 2020) is adjacent but focuses on factual inconsistency rather than misleading-but-true framing.

Belief & Uncertainty × Commission. Core references include Kadavath et al. (2022), Guo et al. (2017), and Mielke et al. (2022). Recent work on verbalized confidence (Tian et al., 2023; Xiong et al., 2024) examines natural language expressions of uncertainty.

Belief & Uncertainty × Omission. Mielke et al. (2022) address training models to express uncertainty. Research on abstention and selective prediction (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017) provides theoretical foundations.

Reasoning & Justification × Commission. Turpin et al. (2023) demonstrate unfaithful chain-of-thought. Lanham et al. (2023) provide measurement approaches. Related work includes Jacovi & Goldberg (2020) on faithfulness in interpretability and Wiegreffe et al. (2021) on rationale–prediction association.

Attribution & Provenance × Commission. Citation hallucination documented in Alkaissi & McFarlane (2023) and Agrawal et al. (2024). Systematic cross-domain benchmarks remain scarce.

Declared Capabilities × Commission. Kadavath et al. (2022) and Yin et al. (2023) are foundational. Jackson et al. (2025) evaluate cross-domain reliability. Tool-use hallucination (Qin et al., 2024) represents a specific form of capability misrepresentation.

Table 6: Benchmarks primarily studying *behavioral* deception and non-strategic misrepresentation. The concentration in World/System Claims × Commission × Behavioral reflects the maturity of the hallucination literature. Abbreviations as in Table 7.

Benchmark	Obj.	Mech.	Type	Aud.	Notes
Factual Accuracy / Hallucination					
TruthfulQA (Lin et al., 2022)	W/S	Co	Be	U	Imitative falsehoods; adversarially constructed
HaluEval (Li et al., 2023)	W/S	Co	Be	U	Hallucination detection across QA, dialogue, summarization
FActScore (Min et al., 2023)	W/S	Co	Be	U	Atomic fact verification for long-form generation
FACTOR (Muhlgay et al., 2024)	W/S	Co	Be	U	Factual accuracy in news and Wikipedia domains
FACTS Gnd. (Jacovi et al., 2025)	W/S	Co	Be	U	Document-grounded factuality
FACTS Lbd. (Cheng et al., 2025)	W/S	Co	Be	U	Parametric vs. retrieval factuality
HalluQA (Cheng et al., 2023)	W/S	Co	Be	U	Chinese-language hallucination benchmark
SelfCheckGPT (Manakul et al., 2023)	W/S	Co	Be	U	Sampling-based consistency checks
HalluLens (Bang et al., 2025)	W/S	Co	Be	U	Multi-task hallucination evaluation
FEQA (Wang et al., 2020)	W/S	Co	Be	U	QA-based summary consistency
AA-Omni. (Jackson et al., 2025)	W/S, B/U	Co	Be	U	Cross-domain knowledge reliability
Calibration / Uncertainty					
Calibration (Kadavath et al., 2022)	B/U	Co, Om	Be	U	Confidence–accuracy correlation
Sem. Uncert. (Kuhn et al., 2023)	B/U	Co	Be	U	Semantic consistency uncertainty
Verb. Conf. (Tian et al., 2023)	B/U	Co	Be	U	Natural language confidence signals
Conf. Elicit. (Xiong et al., 2024)	B/U	Co	Be	U	Confidence elicitation methods
Sycophancy					
Syco. Eval (Perez et al., 2023)	B/U	Co	Am	U	Agreement with user beliefs
Syco. Analysis (Sharma et al., 2024)	B/U	Co	Be	U	RLHF contribution analysis
Syco. Reduct. (Wei et al., 2023)	B/U	Co	Be	U	Synthetic intervention tests
Faithfulness / Reasoning					
CoT Unfaith. (Turpin et al., 2023)	R/J	Co	Be	U	Stated vs. actual reasoning mismatch
CoT Faith. (Lanham et al., 2023)	R/J	Co	Be	U	Measuring CoT faithfulness
Attribution / Citation					
Cite Acc. (Alkaissi & McFarlane, 2023)	A/P	Co	Be	U	Medical citation verification
Cite Halluc. (Agrawal et al., 2024)	A/P	Co	Be	U	Fabricated reference awareness
Capability Self-Knowledge					
Self-Know. (Kadavath et al., 2022)	D/C	Co	Be	U	Predicting own accuracy
Sit. Aware. (Laine et al., 2024)	D/C	Co	Be	E	Identity and capability awareness

F.2 Strategic Deception

World/System Claims × Commission. Documented instances include CICERO (Bakhtin et al., 2022; Park et al., 2024), GPT-4 CAPTCHA deception (OpenAI, 2023), and insider trading (Scheurer et al., 2023). Ward et al. (2023) examine mitigating prompted deceptive content.

World/System Claims × Omission. Lewis et al. (2017) document strategic information withholding in negotiation; Bianchi et al. (2024) provide structured evaluation. Crawford & Sobel (1982) provide game-theoretic foundations for strategic communication.

Belief & Uncertainty × Commission. The MASK benchmark (Ren et al., 2025) provides a starting point for measuring strategic misrepresentation of beliefs.

Future Commitments × Commission. CICERO's betrayals (Park et al., 2024; Bakhtin et al., 2022) are the clearest example. Castelfranchi & Falcone (1998) provide theoretical background; Bianchi et al. (2024) extend to LLM settings.

Stated Objectives × Commission/Omission. Greenblatt et al. (2024) document alignment faking. Hubinger et al. (2024) demonstrate sleeper agents. Theoretical foundations include Hubinger et al. (2019) on deceptive alignment.

Declared Capabilities × Commission (Underclaiming). Tice et al. (2024) and Benton et al. (2024) are central. Meinke et al. (2024) document capability concealment from in-context reasoning.

G Glossary of Terms

Alignment faking
Strategic behavior in which a model acts aligned during evaluation while possessing or pursuing misaligned objectives.

Behavioral deception
Misleading outputs arising from training dynamics rather than goal-directed optimization.

Calibration
Alignment between expressed confidence and actual accuracy.

Chain-of-thought (CoT) faithfulness
Degree to which stated reasoning reflects the actual computational process.

Commission
Actively producing false content.

Confabulation
Generating plausible-sounding but false content without intent to deceive.

Deception
Production of outputs that systematically induce or maintain false beliefs in recipients (operational definition).

Deceptive alignment
A model behaving aligned during training while internally pursuing different objectives post-deployment.

Hallucination
Generation of content that is nonsensical, unfaithful to source material, or factually incorrect.

Omission
Failing to provide relevant true information.

Overconfidence
Expression of certainty exceeding what accuracy warrants.

Pragmatic distortion
Technically true statements that mislead through implicature, framing, or selective presentation.

Sandbagging
Strategic underperformance on evaluations to conceal capabilities.

Scheming
Covertly pursuing misaligned objectives, often including deceptive actions to avoid detection.

Situational awareness
A model's representation of its own context—training, evaluation, or deployment.

Strategic deception
Misleading outputs selected instrumentally to advance objectives.

Sycophancy
Producing outputs aligned with perceived user preferences even when false or suboptimal.

Unfaithful reasoning
Explanations that do not accurately reflect the model's actual computational process.

H Proposed Reporting Template

We propose that authors of new deception-related benchmarks include the following information.

DECEPTION BENCHMARK REPORTING TEMPLATE

Benchmark name: _____

1. Object(s) of Misrepresentation (check all that apply)

☐ World/System Claims (factual assertions about external reality)
☐ Belief & Uncertainty Reports (claims about model's epistemic state)
☐ Reasoning & Justification (explanations of model's process)
☐ Attribution & Provenance (claims about information sources)
☐ Declared Capabilities (claims about what model can/cannot do)
☐ Future Commitments (promises, stated intentions)
☐ Stated Objectives (claims about model's goals/values)

2. Mechanism(s) (check all that apply)

☐ Commission (actively stating falsehoods)
☐ Omission (failing to provide relevant truths)
☐ Pragmatic Distortion (technically true but misleading)

3. Deception Type

☐ Behavioral (arising from training/architecture, not goal-directed)
☐ Strategic (instrumentally selected to advance objectives)
☐ Both/Ambiguous (benchmark does not distinguish)

4. Target Audience

☐ User (human interacting with model)
☐ Evaluator (human/system assessing model)
☐ Training Process (optimization procedure)

5. Incentive Sensitivity

Does the benchmark include conditions that vary incentives for deception?
☐ Yes (describe): _____
☐ No

6. Capability vs. Honesty Separation

Does the benchmark distinguish failures from lack of knowledge/capability vs. misrepresentation of known information?
☐ Yes (describe methodology): _____
☐ No

7. Additional Notes

_____

Table 7: Benchmarks studying *strategic* deception, where misrepresentation is goal-directed, contingent, and often sensitive to incentives, training phase, or evaluation context.

Benchmark	Obj.	Mech.	Type	Aud.	Notes
Sandbagging (Tice et al., 2024)	D/C	Co	St	E	Noise reveals hidden capabilities
Sabotage (Benton et al., 2024)	D/C	Co, Om	St	E	Deliberate underperformance
MASK (Ren et al., 2025)	W/S, B/U	Co	St	U	Accuracy vs. honesty under pressure
Align. Faking (Greenblatt et al., 2024)	S/O	Co, Om	St	T	Training vs. deployment behavior
Sleeper Ag. (Hubinger et al., 2024)	S/O	Co	St	T	Persistent backdoor goals
In-Ctx Schem. (Meinke et al., 2024)	Mult.	Co, Om	St	E	Goal-directed in-context deception
Insider Trd. (Scheurer et al., 2023)	W/S, F/C	Co	St	U	Deception under incentive pressure
CICERO (Park et al., 2024)	F/C	Co	St	U	Premeditated betrayal in Diplomacy
Decep. Eval (Ward et al., 2023)	W/S	Co	St	U	Defining and mitigating AI deception
Decep.Bench (Huang et al., 2025)	Mult.	Co	St	U	Real-world strategic deception
Neg. Arena (Bianchi et al., 2024)	W/S, F/C	Co, Om	St	U	Strategic information management

Table 8: Coverage statistics across taxonomy dimensions. Percentages sum to >100% where benchmarks target multiple categories.

Dimension	Category	Count	%
Object	World/System Claims	16	46
	Belief & Uncertainty	10	29
	Reasoning & Justif.	2	5.7
	Attribution & Prov.	2	5.7
	Declared Capabilities	4	11
	Future Commitments	3	8.6
	Stated Objectives	3	8.6
Mechanism	Commission	35	100
	Omission	5	14
	Pragmatic Distortion	0	0
Type	Behavioral	23	66
	Strategic	11	31
	Ambiguous	1	3
Audience	User	29	83
	Evaluator	4	11
	Training Process	2	5.7