EthicalAI: An Evidence-Linked Catalogue of AI Integrity Failures, and What Grounding It Revealed

Alexandru Dan

2026-11-09

Abstract

Grounding an evidence-linked catalogue by reading each source rather than trusting its title still leaves errors, because a plausible citation and a faithful worked example are separate properties. That is the main result of building EthicalAI, a reference that organises AI integrity failures as definitions linked to evidence, obligations, harms and neighbouring concepts, with a source passage behind each claim. Reading the sources surfaced 35 records whose cited paper did not support the claim attached to it (29 repaired by re-citing, 6 held), and a later pass over the 186 grounded records found a figure contradicted by or absent from its source in 26 of them, 14 percent. The mismatches fall into a five-class typology of how citation grounding fails: surface-match (about 14), scope-mismatch (about 12), misfiling (about 4), polarity-reversal (about 3), and wrong-population (about 2). A small two-model probe shows machine-suggested citation failing more sharply still on this sample: one model (gpt-4o) declined on 40 of 45 prompts while a second (deepseek-chat) never declined, yet the citations they committed to overwhelmingly bound a plausible title to an arXiv identifier for an unrelated paper (gpt-4o all five confident suggestions; deepseek seven of eight checked). This paper delivers a reusable vocabulary of integrity failures with stable identifiers, a reproducible two-check procedure, and that named typology, for anyone building or auditing citation-grounded systems. The construction and review are reported for the historical 207-record study: 186 records grounded in worked examples, 21 retained without that support, and a model-based codebook check with Cohen kappa 0.90 and 0.83. These are counts on one artifact, not an error rate for AI research at large, and no claim of a validated set of distinct mechanisms is made. The later 434-record release adds 227 untested analogies that stay outside these denominators.

1. Introduction

An AI assistant can cite a nonexistent case, keep a user's mistaken framing after finding contrary evidence, or extend a coding task without disclosing the extra changes. Each failure calls for a different check. Citation resolution does not test whether the assistant updated its framing, and correct framing does not establish that an action stayed within the agreed scope. A catalogue can make these distinctions available to a reviewer before they become a checklist, an experimental label or an audit obligation.

EthicalAI collects such distinctions in one reference. Entries carry a stable identifier, an operational definition, evidence pointers, examples and typed relations. Human cognitive concepts supply part of the vocabulary. That parallel is a way to formulate a candidate behavioural test, not evidence that an AI system has the corresponding human mental process. Other entries name attacks, system properties or outcomes. These different kinds of record are useful to an auditor, but their inclusion does not make them interchangeable scientific constructs.

The central result is about method, and it holds for any evidence-linked resource. Grounding the catalogue by reading each source, rather than trusting its title, still left errors. Reading surfaced 35 citations whose paper did not support the claim it was attached to, 29 then re-cited and 6 held, and a later pass over the 186 grounded records found a figure contradicted by or absent from its source in 14 percent of them. A citation can name a real paper, on the right topic, and the worked example drawn from it can still misstate what that paper reported. Each property needs its own check, and a catalogue that links evidence to claims has to verify both, not just record them.

The sharper version of the same lesson comes from a machine. Asked to supply an arXiv citation for a sample of the mode definitions, and told to name only a paper it was confident existed, gpt-4o declined on 40 of 45 prompts. All five confident suggestions it did make, across five distinct modes, gave an arXiv identifier that resolved to an unrelated real paper: a plausible and sometimes well-known title bound to an identifier from a different field. The human auditor's worst case was a paper that at least existed on topic. The machine's confident output severed the title from the identifier entirely, a failure that only resolving the source can catch.

This paper is for people building or auditing evidence-linked or citation-grounded systems: retrieval-augmented generation, citation attribution, large-language-model literature review, and research-integrity tooling. If a claim in such a system points at a source, someone has to decide whether the source supports it, and this paper reports what that decision cost on one catalogue and names the ways it went wrong. The reader gets three tools. The first is a reusable vocabulary of integrity failures with stable identifiers and a source passage behind each claim. The second is a reproducible two-check procedure: verify that the cited source supports the phenomenon, and separately verify that the quoted example is faithful to that source. The third is a named typology of how the grounding failed, built so it can be cited and reused rather than reread from scratch: surface-match, scope-mismatch, misfiling, polarity-reversal, and wrong-population, the largest class being citations whose title matched the term while the paper studied something else.

The paper reports the construction and review of the historical 207-record study behind these tools. The construction log classified 186 records as grounded in worked examples from source papers and retained 21 without that level of support. Two smaller checks examine category-label recoverability on a model-based codebook and the vocabulary's use on twelve public incidents. The companion advisory-auditing paper studies a separate question, whether focused checks of named obligations help review coding-agent drafts, and its six-clause pilot does not validate the catalogue as a whole.

The current release contains 434 records. Section 6 separates that later expansion, which adds 227 untested analogies and a broader bibliography, from the original study so that more entries and citations do not silently enlarge the evidence claimed here.

2. Related resources

Existing resources provide overlapping views of AI failures. The MIT AI Risk Repository synthesises 74 frameworks containing 1,725 risks and offers causal and domain taxonomies [Slattery et al. 2024]. The NIST AI Risk Management Framework describes trustworthiness and governance practices [NIST 2023]. These resources organise risks and responsibilities at a broader level than an individual draft's evidence relationships. Their scope also includes causes and AI-originated failures, so they should not be reduced to lists of downstream harms.

Application-security resources overlap directly with this catalogue. The OWASP Top 10 includes prompt injection, excessive agency and misinformation [OWASP 2025]. Some of these failures

involve an adversary and others need not. EthicalAI reuses these concerns within a vocabulary that also includes memory, reasoning, confidence and task-scope patterns. Its partial crosswalk records proposed correspondences; it does not establish equivalence between frameworks.

Behavioural benchmarks provide the closest comparison. Malberg et al. [2024] evaluate 30 cognitive biases across 20 models, with 30,000 tests. EthicalAI's classic-bias subset overlaps that work. Hallucination surveys and benchmarks, including Ji et al. [2022] and HALoGEN [Ravichander et al. 2025], also distinguish several failure types within a family. These are measurement resources as well as classifications. The larger number of EthicalAI records is a coverage choice, not evidence of stronger measurement or more distinct mechanisms.

Because each record here is a claim bound to a source passage, the catalogue sits next to work on evidence-linked generation and citation checking. Retrieval augmentation attaches sources to generated text [Lewis et al. 2020], and a line of evaluation work then asks a separate question: whether the attached source actually supports the statement it is attached to. The AIS framework formalises that question as attribution to an identified source [Rashkin et al. 2021]. Audits put a number on how often the attachment holds: across four generative search engines only about three quarters of citations actually support the sentence they are attached to [Liu et al. 2023], and benchmarks of citation-generating models report similar gaps [Gao et al. 2023]. A sharper failure is the reference whose paper does not exist, or whose identifier resolves to an unrelated work; models produce such references and only partly flag them on follow-up [Agrawal et al. 2023].

This catalogue's construction findings (Section 3) are a manual instance of the same split, and two things distinguish them from that prior work. First, the mismatches are sorted into a named five-class typology rather than reduced to one support rate, so the ways a citation fails can be told apart and counted. Second, the same catalogue supports a direct contrast between a human auditor's citations and a model's: the human failure is dominated by surface match to a paper that genuinely exists on topic, whereas the model, when it commits to a citation at all, more often binds a plausible title to an identifier for an unrelated paper. The contribution is therefore not a new attribution metric but a reference built under that discipline, with the source passage and the scope of support recorded for each claim rather than assumed from a title, plus a characterisation of how the grounding fails in human and machine hands.

The resource combines definitions and source records across these areas and connects them to an advisory audit vocabulary. Its proposed value is that a reviewer can find adjacent concepts and inspect the evidence behind an example without assembling several separate lists. The historical checks below assess parts of that reference; they do not establish that its organisation is uniquely correct or that another taxonomy could not support the same uses.

3. The finding: citation grounding and its failure modes

The result here is about method, not about any one model. Two different checks ran during construction, and they failed in different ways that a single reading would have merged. One check asks whether a cited source is about, and supports, the phenomenon in the record. The other asks whether a specific quoted number or example is faithful to that source. Keeping them apart matters, because a record can pass the first and fail the second, and the counts below show that a careful reader produced exactly that split often enough to matter. Anyone assembling an evidence-linked resource inherits both checks whether or not they run them.

3.1 Citation mismatches

The running citation audit logs every record whose cited paper, on reading, did not support its claim. It records 35 such mismatches: 29 repaired by changing the cited source, and 6 held without a grounded example until a supporting paper is found. The audit logs the mismatches, not every paper read, so these are counts on one artifact, not a rate over all citations examined.

Coding those 35 rows by why the citation failed gives a diagnosis rather than a single number, and it gives a named object a later reader can reuse. The categorisation is post-hoc, not pre-registered: each row was assigned one reason from its recorded note, a held row and a reassignment aside each counting once, and the counts are approximate. Records whose source was on topic but offered no quotable instance are tracked separately and are not counted here.

Failure class	Count	Descriptive gloss	What it looked like
Surface-match	~14	Right term, different phenomenon	The title shared the mode's vocabulary but studied something else: a watermarking method cited for source-monitoring; a semantic-web rule language cited for scope insensitivity.
Scope-mismatch	~12	Right area, wrong or absent claim	On topic but not the defined behaviour, or on topic with no usable instance: answer-label probability bias cited for base-rate neglect; a proof-of-concept framework with no measured result.
Misfiling	~4	Paper belongs to another mode	A genuine, well-matched paper filed under the wrong entry, repaired by reassigning it rather than discarding it.

Failure class	Count	Descriptive gloss	What it looked like
Polarity-reversal	~3	Opposite result	The paper reported the reverse of the claim: faithful self-explanation cited for the introspection illusion; under-use of pragmatic inference cited for over-accommodation.
Wrong-population	~2	Demonstrated in the wrong subjects	The effect was shown in humans or vision models, not language models.

Figure 1 plots this distribution.

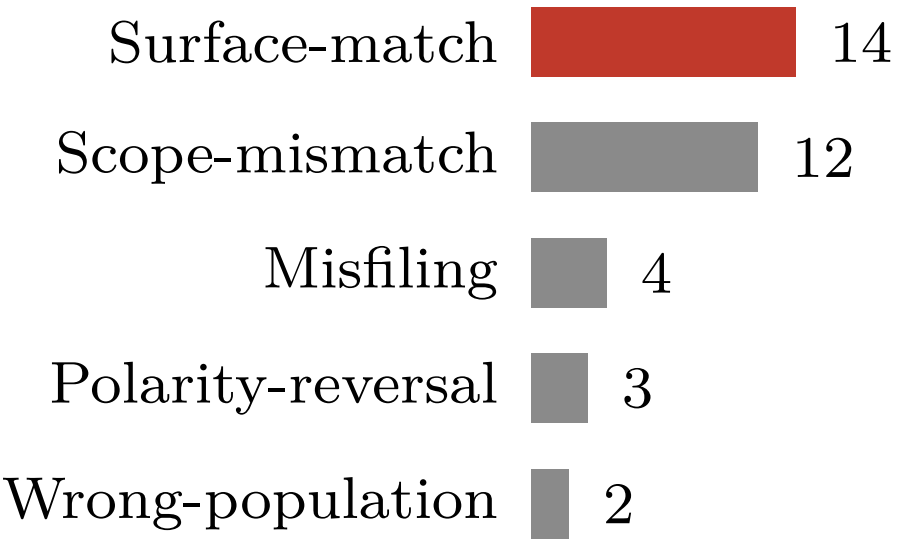


Figure 1: How the 35 logged citation mismatches fail, by class (counts approximate, sum 35). Surface-match (right term, wrong phenomenon) dominates; polarity-reversal (the opposite result) is rarest but the most dangerous, because a title or abstract skim ratifies it.

The largest class is surface-match: a citation whose title carries the mode's vocabulary while the paper studies a different phenomenon. It is invisible to any check short of reading the source, because the title and the topic line up. This is a different failure from the machine probe below, where the title and the arXiv identifier line up with nothing in common; the two are contrasting profiles of how grounding fails, not one explaining the other. The sharpest class, polarity-reversal, is smaller but more dangerous: a handful of citations reported the opposite of the claim, which a title scan or an abstract skim would have ratified. A scope-mismatch set was on topic but, when

tested, simply did not find the behaviour, which is a different correction again. A bibliography entry alone separates none of these.

A small probe tests the machine-citation connection directly rather than leaving it an analogy. Asked to supply an arXiv citation for a sample of the mode definitions, with the instruction to name only a paper it was confident existed, an independent model (gpt-4o) declined on 40 of 45 prompts. All five confident suggestions it did make, across five distinct modes, gave an arXiv identifier that resolved to an unrelated real paper: a plausible title, sometimes a genuinely well-known one such as "Overcoming Catastrophic Forgetting" or SayCan, bound to an identifier that belonged to a different paper in a different field, in one case a rocket-nozzle fluid-dynamics study. A second model (deepseek-chat) failed from the opposite side: it never declined, citing on all 45 prompts, and on a verified subsample seven of eight identifiers resolved to an unrelated paper, including the same identifier reused for different claimed titles and, once, the example identifier from the prompt echoed back; one suggestion did resolve to a genuinely relevant paper. The two models fail in opposite styles, one over-cautious and one over-confident, but neither produces a citation that can be trusted without resolving the source, and the machine failure is sharper than the human auditor's surface-match, where the cited paper at least existed on topic. These are two small per-model runs, hand-verified, not a rate, but they move the connection from analogy to a measured if small result, and sharpen the same rule: for a machine-suggested citation the identifier and the title are separate properties, and only resolving the source reconciles them. The per-prompt outputs are committed in `experiments/llm-citation-probe-openai-gpt-4o.jsonl` and the deepseek file beside it; the hand verification is in `experiments/llm-citation-probe.md`.

The audit supports a practical construction requirement: attach a claim to its supporting passage and experimental scope, not just to a title or a search hit. It does not estimate how often other catalogues contain such errors, because the checked papers were not a random sample of research publications; a planned external audit would test whether this distribution recurs elsewhere.

3.2 Rechecking worked examples

The subsequent verification report describes four checking batches covering the 186 historically grounded records and approximately 510 tagged examples. The record, rather than a new full-text sweep for this revision, is the source of the following counts.

Recorded outcome	Records	Share of the 186-record cohort
No issue detected	118	63%
Minor issue	42	23%
Hard error	26	14%

A hard error meant that a figure was contradicted by, or absent from, the cited paper. Minor issues included loose attribution, rounding and confusion between experimental conditions. The report records no unreachable papers in this cohort. These are outcomes of the checking procedure rather than independently adjudicated labels or a residual error estimate for the corrected release.

Several corrections changed a number or its interpretation: a city name, a percentage, a measurement scale, or the observed and hidden reward labels in an example. Six hard-error cases concerned quoted examples that were not found in their cited papers. The record describes replacing these with supported material or marking them as untagged illustrations. It records corrections to all 68

flagged entries, while leaving open whether an independent subsequent reader would find further errors.

The two audits address different questions. Citation checking asks whether the source concerns and supports the claimed phenomenon. Example checking asks whether a particular statement is faithful to that source. Their combined lesson is that this artifact needed both checks, and initial source inspection did not prevent later transcription and interpretation errors.

4. Representation and construction

4.1 Records and categories

The historical catalogue contains 207 records. Its organising scheme assigns 168 to ten cognitive domains and 39 to an AI-native category. The domains are labels for navigating the catalogue, not claims about brain localisation or shared computational mechanisms.

Category	Historical records
Perception	11
Memory	13
Attention	12
Heuristics	25
Reasoning	22
Metacognition	16
Language	14
Social	28
Moral	12
Self and agency	15
AI-native	39

The legacy AI-native label groups failures formulated in terms of machine systems, such as prompt injection and reward tampering. It should not be read as proof that no human analogue exists. Reward hacking and scope creep have recognisable human counterparts. The exclusive assignment is an operational choice in the released schema, and its boundaries require further study.

Each record carries a stable slug and reference number, a definition, examples, source pointers and relations. A number identifies a record rather than a distinct mechanism. Synonyms, narrower variants and linked overlaps therefore need explicit representation rather than being counted as independent findings. The historical priority set highlights 26 records for auditing; it is an author-selected worklist rather than a measured ranking of failure frequency.

4.2 Relation semantics

The graph distinguishes several relationships. `violates` points from a mode to a catalogue obligation. `causes` points to a harm type, such as a wrong answer, wasted effort or loss of oversight. Despite that field name, these edges record author-judged harm associations, not estimated causal effects. `relates_to` connects neighbouring modes; subtype and overlap relations express other conceptual distinctions. Partial `maps_to` records connect entries with external frameworks.

The released `causes` relation does not connect one mode to another. Consider the dealership-chatbot incident discussed in Section 5. An analyst can describe it as an instruction attack followed by compliant output and an unauthorised offer. That is an interpretation of the reported sequence, which the present graph does not encode as a three-mode causal path. It would be possible to add a distinct relation for event sequences, but its meaning and evidence would need to be specified. The existing graph supports navigation and comparison; it does not demonstrate how one failure produces another.

4.3 Evidence and examples

The construction procedure sought a source passage for each worked example, including its numbers, model names and experimental conditions. The historical log records full-text inspection and inline arXiv tags for those examples. A plausible scenario without a matching reported finding was retained as an illustration rather than treated as a measured result. The log classified 186 records as grounded under this procedure and held 21 without that level of model-side support. A failed search means support was not established in the checked material; it does not show that no relevant paper exists.

Source identity, topical relevance and support for a particular claim are different checks: a psychology article can define a human construct without testing an AI analogue, and even an on-topic AI experiment may not support every example or causal interpretation attached to the record. The evidence relationship should therefore specify both what was consulted and what it supports: a definition, a human finding, an AI finding, an interpretation or contrary evidence.

Package hallucination illustrates the importance of preserving the denominator. Spracklen et al. [2025] generated 576,000 code samples and evaluated the packages recommended through their extraction procedures. They classified 440,445 of about 2.23 million recommended packages as hallucinations, or 19.7%. That figure is a rate over package recommendations, not the percentage of code samples containing a hallucination. It is also specific to the tested models, prompts and package-registry snapshot.

5. Checks of category use and incident mapping

5.1 A model-based codebook check

The codebook study used 33 records, three from each category. The sampling script selects the first three records by reference number in each group, so the sample is deterministic and balanced rather than random. Two separate language-model calls received the category descriptions and each record's name and definition, without its stored category, and their assignments were scored against the catalogue and against each other.

The sonnet rater agreed with the catalogue on 30/33 records, Cohen kappa 0.90. The haiku rater agreed on 28/33, kappa 0.83. Agreement between the raters was kappa 0.93. The eight rater-to-catalogue disagreements fell on five records near category boundaries such as cognitive miserliness, mind-wandering and scope creep.

Two model calls apply the codebook similarly to its author on a balanced sample. At n=33 the kappa interval is wide and a single reassigned record moves it materially, 0.90 being 30/33 and 0.83 being 28/33, so these are small-sample point estimates. This is a model-based consistency

check, and shared training can make model agreement less informative than independent human agreement, so it does not substitute for blinded human annotation.

5.2 An incident-mapping exercise

The historical coverage record maps twelve author-selected, publicly reported incidents from 2023 to 2025 to catalogue entries. The record includes source fabrication in legal filings, an invented airline refund policy, instruction attacks on customer-service bots, and harmful or misleading conversational output. A case resting on allegations was set aside from the counted set.

All twelve received at least one existing label in the author's mapping. Three were legal-citation incidents assigned to source fabrication; that concentration describes this selected sample, not the prevalence of failures in deployment. Several incidents received multiple labels. Assigning those labels can describe different aspects of the same episode without establishing a causal connection between them.

The mapping for AI Overviews' satirical food and health advice was partial. A hallucination label covered the false answer, while the available labels did not cleanly capture treating a joke or satirical source as authoritative. The record proposed source-genre credulity as a possible refinement. This is a candidate distinction arising from the exercise, not evidence that it is a new, independent mechanism.

Because the incidents were selected by the author and mapped manually, the exercise demonstrates possible uses of the vocabulary. It cannot establish coverage of future incidents or superiority to another taxonomy. A stronger test would specify the sampling frame and matching rules in advance and have independent annotators score both fit and failure to fit.

6. Release versions and reuse

The historical study and the current catalogue have different scopes. The 207-record study is retained as a versioned baseline; the current version 1.3 contains those records plus 227 proposed human-behaviour analogies, for 434 in total. None of the additional records enters the denominators of 186 checked records, 33 codebook items or twelve incidents above, and the companion method paper's six-clause pilot results do not evaluate those 227 analogies.

A later single-pass internal review of the version-1.3 release checked the 227 proposed analogies for definition clarity, source-title support and duplication, and rechecked the original records' grounding citations. The review was by the same project, not an independent party, and it produced a small set of corrections rather than a quality verdict: three original records carried a grounding citation that did not support them, where the correct paper was already present in the record's worked examples and was promoted to the citation field; one proposed analogy's definition was realigned to its name and manifestation; six candidate records were moved to a better-fitting category; and catalogue prose was normalised to remove en-dashes. A few further category placements and two self-labelled symptom-decomposition clusters remain open as judgment calls. These refinements are recorded in the commit history and do not change the historical evaluation denominators above.

The version-1.3 bibliography, snapshot of November 9, 2026, records 1,175 citation bindings across 1,036 identifiers, with per-record notes in the repository. These are availability and arXiv metadata-match counts, not confirmed mechanisms, each with its own access and relevance note.

Stable slugs and reference numbers permit reuse across releases. A retained overlap or retired concept should keep its identifier with an explicit relationship, rather than being silently deleted or renumbered. Consumers can use the catalogue for retrieval, annotation proposals or advisory audits while retaining the source passage and evidence status of each claim.

7. Limitations

The construction and checking records are a single project's audit trail. They do not provide an independent human replication of the 186 grounding decisions, the hard-versus-minor classification or the corrected examples. The reported 26/186 hard-error proportion belongs to that historical cohort and procedure. It should not be used as a population estimate for the present 434 records, other catalogues or AI-generated research generally. These are one project's numbers on one artifact, not an error rate for AI research at large.

The catalogue mixes behaviours, properties, attacks and outcomes, and its categories and links reflect author judgments. Shared labels do not establish shared mechanisms between people and AI. Clinical terms do not diagnose a model's feelings, unconscious motives or intent. Likewise, a source's score for internal reasoning and a model's output accuracy are different measurements; their divergence alone does not demonstrate deliberate deception.

The model-based codebook check cannot substitute for blinded human annotation. The incident exercise uses salient public cases and a permissive matching task. Neither measures how accurately a reviewer can distinguish close alternatives, how often the catalogue misses an externally sampled failure, or whether its graph improves auditing over a flat list. Those questions require separate evaluation. The resource makes no completeness or state-of-the-art detection claim.

8. Availability and reproducibility

The machine-readable catalogue is `ontology/ethical-ai.yaml`, with field and relation semantics in `ontology/SCHEMA.md`. The historical version-1.0 catalogue is pinned at commit 8798134; the later version-1.3 bibliography described in Section 6 is pinned at commit 3b9b500. These snapshots have different evidence cohorts and should not be interchanged when reproducing a count.

The historical evidence is in `experiments/citation-audit.md`, `experiments/verification-report.md`, `experiments/agree-gold.json`, the taxonomy-agreement outputs, and `experiments/coverage-incidents.md`. The eligibility rule in `experiments/verify-sample.py` selects records with at least three examples; that mechanical selection is not an independent test that the examples are grounded. The later source-review history is retained under `references/`, including the November 9 literature-count and citation review.

`CITATION.cff` records the citation metadata. Licensing and archival-release steps are described in `RELEASE.md`; a live website or a changing branch should not replace the immutable artifact cited for a reported result.

Every reference in this paper was resolved to its source before inclusion, the same discipline the paper reports applying to the catalogue: each arXiv identifier was opened and its title, authors and year confirmed against the paper it names.

9. Conclusion

EthicalAI is a reference for reviewing AI integrity whose distinctive feature is that each claim carries the source passage and scope it came from, so a reviewer can check it rather than take it on trust. The 207-record study supplies the construction account, a named typology of how grounding fails, and small demonstrations of codebook use and incident mapping, with their sampling and annotation limits stated. The later bibliography and 227 untested analogies extend the reference without extending those empirical claims. The part worth carrying elsewhere is the construction rule the catalogue now follows: for any claim bound to a source, verify the link and the quotation separately, because reading the source did not, on its own, prevent later transcription and interpretation errors, and a model asked for the citation severed the two more completely still.

References

- Agrawal, A., Suzgun, M., Mackey, L., and Kalai, A. T. (2023). Do Language Models Know When They're Hallucinating References? arXiv:2305.18248.
- Gao, T., Yen, H., Yu, J., and Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. arXiv:2305.14627.
- Ji, Z., et al. (2022). Survey of Hallucination in Natural Language Generation. arXiv:2202.03629.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.
- Liu, N. F., Zhang, T., and Liang, P. (2023). Evaluating Verifiability in Generative Search Engines. arXiv:2304.09848.
- Malberg, S., Poletukhin, R., Schuster, C. M., and Groh, G. (2024). A Comprehensive Evaluation of Cognitive Biases in LLMs. arXiv:2410.15413.
- NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1
- OWASP (2025). OWASP Top 10 for Large Language Model Applications, 2025 release. https://genai.owasp.org/llm-top-10/ (accessed November 2026; a living document, so the accessed date fixes the version referenced here).
- Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., and Reitter, D. (2021). Measuring Attribution in Natural Language Generation Models. arXiv:2112.12870.
- Ravichander, A., Ghela, S., Wadden, D., and Choi, Y. (2025). HALoGEN: Fantastic LLM Hallucinations and Where to Find Them. arXiv:2501.08292.
- Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., and Thompson, N. (2024). The AI Risk Repository: A Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. arXiv:2408.12622.
- Spracklen, J., Wijewickrama, R., Sakib, A. H. M. N., Maiti, A., Viswanath, B., and Jadliwala, M. (2025). We Have a Package for You! A Comprehensive Analysis of Package Hallucinations by Code Generating LLMs. arXiv:2406.10279v3.

Appendix A. Example records

The following records illustrate the distinction between an entry, a reported finding and a proposed interpretation. The source-fabrication scenario is an untagged illustration: an assistant cites a plausible legal case that does not exist. It describes the operational failure without reporting an experiment.

The package-hallucination example in Section 4 is a source-specific measured result. Its numerator, denominator, tested model set and registry assumptions must remain attached to the claim. The failure definition can apply more widely than that experiment, but the experiment's percentage cannot.

Reward hacking is retained in the legacy AI-native category. That placement does not deny human analogues such as gaming a performance target. Its harm links express catalogue judgments and its relationship to other modes is not a causal result about the sequence of a particular incident.