Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.089115

REVIEW

Accountable NLP for Evidence-Grounded Decision Briefings: A Critical Review and Evaluation Framework

Jihoon Moon*

Department of Data Science, Duksung Women's University, Seoul, Republic of Korea

*Corresponding Author: Jihoon Moon. Email: jmoon25@duksung.ac.kr

Received: 14 July 2026; Accepted: 17 August 2026; Published: 15 September 2026

ABSTRACT: Large language models and retrieval-augmented generation (RAG) systems are increasingly employed to transform evidence into decision-facing briefings, alerts, and recommendations. In these settings, explainability cannot be evaluated merely by fluency, readability, or factual correctness. A briefing may be factually correct while still being unsafe if it cites sources that do not substantiate the claim, suppresses uncertainty, converts correlational evidence into causal language, recommends an unauthorized action, or leaves no auditable path for human review. This review synthesizes 104 sources spanning explainable natural language processing (NLP), faithful explanation, hallucination and factuality evaluation, RAG, citation faithfulness, uncertainty communication, causal language, human–AI interaction, engineering and regulatory decision support, and institutional accountability. It makes four contributions. First, it defines evidence-grounded decision briefings as a distinct NLP setting characterized by identifiable evidence inputs, constrained decision-facing outputs, and minimum accountability requirements. Second, it proposes a five-layer taxonomy encompassing evidence representation, explanation generation, retrieval and source grounding, verification and evaluation, and human accountability. Third, it develops an operational evaluation framework for claims, citations, uncertainty statements, causal wording, action labels, and complete briefing episodes. Fourth, it complements the conceptual synthesis with source-level trend analyses, targeted quantitative comparisons, and representative use cases drawn from generic, engineering, and regulatory contexts. The synthesis reveals that existing surveys provide critical foundations but do not jointly address five interdependent requirements: citation-to-claim entailment (whether the cited evidence supports the exact claim), causal-language discipline, uncertainty preservation, action appropriateness, and human accountability in decision-facing generated text. This review concludes with open challenges for claim segmentation, retrieval adequacy, citation-to-claim entailment, causal test suites, uncertainty preservation, accountability logging, and preference-bias-aware human evaluation.

KEYWORDS: Explainable natural language processing (NLP); accountable NLP; evidence-grounded generation; retrieval-augmented generation (RAG); faithful explanation; source entailment; hallucination; causal language; action appropriateness; human accountability

1 Introduction

Neural natural language processing (NLP) systems are no longer used only to classify, summarize, translate, or retrieve text. They increasingly serve as language interfaces between complex evidence sources and human decision-makers. A generated briefing may compress retrieved policy passages, tables, model predictions, uncertainty intervals, feature attributions, and operator context into a short explanation or recommendation. This shift in usage creates a form of explainability unlike conventional post hoc interpretation: the explanation is a generated textual artifact that may influence downstream judgment. The

problem is timely because retrieval-augmented generation (RAG), long-context models, and instruction-tuned language models are being adopted in domains where briefings are expected to support triage, review, escalation, compliance, or risk communication. In such settings, ordinary quality criteria are insufficient. A briefing may be fluent, readable, factually correct, and well cited yet still be unsupported by the cited evidence, overconfident, or operationally unsafe. Therefore, accountable NLP requires a richer evaluation vocabulary than surface fluency or generic factuality.

This review defines the focal setting as an evidence-grounded decision briefing: a generated or semi-generated textual artifact that converts one or more pieces of evidence into decision-facing communication. The evidence may include retrieved passages, structured fields, tabular values, predictions, uncertainty estimates, feature attributions, text spans identified as rationales, rule fragments, or contextual constraints. The output is not merely a summary of evidence; it organizes evidence for a reader who must decide whether to inspect, accept, hold, escalate, or revise a course of action. This review is not motivated by a lack of research on explainable NLP or hallucination. Existing surveys have mapped explainable NLP, RAG, hallucination, and automated fact-checking [1–4]. A related survey tradition examines faithful model explanation and the distinction between plausible and faithful explanations [5]. However, prior surveys do not evaluate decision-facing briefings across source support, uncertainty preservation, causal restraint, action authorization, and human accountability within a unified framework.

Recent reviews published in 2026 have examined large language model (LLM) applications across NLP, architectural evolution, deployment challenges, and intrinsic interpretability [6–8]. Other surveys synthesize the broader LLM lifecycle and multilingual deployment, covering pre-training, post-training, utilization, evaluation, language fairness, and accessibility [9,10]. A trustworthy-RAG survey extends this discussion to reliability, privacy, safety, fairness, explainability, and accountability at the system level [11]. These studies establish a contemporary foundation for model development, evaluation, multilingual use, and trustworthy retrieval. Taken together, these surveys provide a strong foundation for explainable NLP, trustworthy RAG, and LLM evaluation. However, they do not offer an operational framework for decision-facing briefings, where source support, uncertainty, causal restraint, action authorization, and human responsibility must be assessed together. Table 1 highlights this gap by comparing representative surveys across these dimensions. Building on their contributions, the present review integrates them into an evidence-to-action accountability framework.

Table 1: Dimension-level comparison of representative surveys and the present review.

Representative Survey	Explanation Faithfulness	Factuality	Retrieval and Citation	Exact Source Entailment	Uncertainty Preservation	Causal Restraint	Action Authorization	Human Auditability
Danilevsky et al. [1]	✓	△	—	—	△	—	—	△
Gao et al. [2]	—	△	✓	△	△	—	—	△
Ji et al. [3]	—	✓	△	△	△	—	—	—
Lyu et al. [5]	✓	△	—	—	△	—	—	△
Ni et al. [11]	△	△	✓	△	△	—	△	✓
Present review	✓	✓	✓	✓	✓	✓	✓	✓

Note: ✓ indicates explicit and substantive coverage; △ indicates partial or indirect coverage; and — indicates that the dimension is not a central evaluation focus. The coding reflects the primary scope of each survey and should not be interpreted as implying that an unmarked or omitted dimension is entirely absent.

This distinction is substantive rather than terminological. Factuality remains an important supporting criterion, addressing whether a statement is true or substantiated by reliable evidence, but it is not treated as one of the five central evaluation dimensions of the proposed framework. The five central evaluation

dimensions are citation-to-claim entailment, causal-language discipline, uncertainty preservation, action appropriateness, and human accountability. Citation-to-claim entailment assesses whether the cited evidence supports the exact claim. Causal-language discipline evaluates whether the wording overstates the epistemic status of the evidence. Uncertainty preservation assesses whether probabilities, intervals, limitations, and epistemic qualifiers are retained without omission or misleading reframing. Action appropriateness examines whether the generated text recommends a response justified by the evidence and applicable rules. Human accountability concerns whether reviews, overrides, disagreements, and assignments of responsibility are traceable. A system that performs well on one dimension may fail on another.

The contributions of this review are fourfold.

- First, this review defines evidence-grounded decision briefings as a distinct NLP task with identifiable evidence inputs, bounded actions, decision-facing language, and traceable review requirements.
- Second, it develops a five-layer taxonomy linking evidence representation, explanation generation, retrieval and source grounding, verification and evaluation, and human accountability.
- Third, it operationalizes joint accountability through claim-level evaluation units, annotation labels, automated checks, expert-review triggers, and audit-record requirements.
- Fourth, it complements the conceptual synthesis with trend analyses, quantitative comparisons, and representative use cases across generic, engineering, and regulatory settings.

The central synthesis of this review is the unified treatment of five dimensions typically studied separately: citation-to-claim entailment, causal-language discipline, uncertainty preservation, action appropriateness, and human accountability. This integrated framework is crucial for decision-facing text generated from evidence because a briefing can be factually plausible and well cited while still being unsafe if it overstates causality, hides uncertainty, recommends an action that the evidence does not authorize, or lacks traceable oversight.

2 Review Methodology and Source Selection

2.1 Review Type and Rationale

This article applies a targeted narrative review design. A narrative review is appropriate because the reviewed literature is methodologically heterogeneous, encompassing explanation benchmarks, hallucination surveys, fact-checking datasets, RAG architectures, citation-evaluation methods, uncertainty communication, causal-language research, interaction between humans and artificial intelligence (AI), and responsible-AI documentation. These streams do not share a single intervention, outcome measure, or dataset from which pooled quantitative effects could be estimated.

This article applies a targeted narrative review because the relevant literature is methodologically heterogeneous. Table 2 consolidates the search scope, query families, eligibility criteria, and final source set, providing a concise and auditable overview of the review procedure.

2.2 Search Dates, Databases, and Query Families

Candidate sources were identified through searches conducted from January to July 2026, including an update search performed during manuscript revision. The updated search specifically targeted recent work on intrinsic LLM interpretability, faithful explanation, source fidelity, hallucination evaluation, retrieval and citation assessment, uncertainty quantification, LLM-based evaluation, and accountable engineering and regulatory decision support.

Table 2: Summary of source-selection and review procedures.

Item	Specification
Search period	January–July 2026, including an update conducted during manuscript revision.
Search sources	Major academic search engines, publisher databases, repositories, and conference proceedings were searched, including Google Scholar, ACL Anthology, ACM Digital Library, IEEE Xplore, arXiv, and OpenReview. Relevant journals and conferences in NLP, trustworthy AI, human–AI interaction, engineering, and regulatory decision support were also reviewed (e.g., ACL, CHI, FAccT, and leading journals in these fields).
Search themes	Explainable and faithful NLP, hallucination and factuality, RAG and citation support, causal and uncertainty-aware language, human and automated evaluation, accountability, and engineering or regulatory decision support.
Inclusion criteria	Studies were included if they directly contributed to explanation, grounding, claim or citation support, uncertainty or causal communication, human evaluation, accountability, or evidence-based decision support.
Exclusion criteria	Studies focused only on prediction, optimization, control, or domain performance were excluded unless they had a clear connection to NLP-based explanation, textual evidence, uncertainty communication, or accountable decision support.
Final source set	The final set included 104 sources covering the five taxonomy layers, foundational studies, recent 2025–2026 research, and representative engineering and regulatory applications. The aim was a focused conceptual review rather than an exhaustive bibliometric survey.

Note: ACL: Association for Computational Linguistics; ACM: Association for Computing Machinery; AI: Artificial Intelligence; IEEE: Institute of Electrical and Electronics Engineers; LLM: Large Language Model; NLP: Natural Language Processing; RAG: Retrieval-Augmented Generation; QA: Question Answering.

The search covered major scholarly databases, disciplinary repositories, and preprint platforms, including ACL Anthology, ACM Digital Library, IEEE Xplore, arXiv, OpenReview, and Google Scholar, which was used for cross-checking. To broaden disciplinary coverage, representative journal articles and conference proceedings were also reviewed in NLP, machine learning, human–AI interaction, explainable AI, and responsible computing. Studies from major venues, including ACL, NeurIPS, CHI, and FAccT, were considered when they directly addressed explanation faithfulness, evidence grounding, evaluation, accountability, or decision-support applications.

The search strings combined four keyword families: (i) "explainable NLP," "faithful explanation," "rationale," "probing," or "causal language"; (ii) "hallucination," "factuality," "claim-level evaluation," "atomic fact," or "citation faithfulness"; (iii) "RAG," "retrieval-augmented generation," "source grounding," "long-context retrieval," or "attributed question answering"; and (iv) "human accountability," "human–AI interaction," "uncertainty communication," "action recommendation," or "decision support." Query variants also included "citation-to-claim entailment," "citation support," "causal overstatement," "uncertainty preservation," "LLM-as-a-judge," "reinforcement learning from human feedback," and "Constitutional AI".

2.3 Inclusion, Exclusion, Screening, and Sufficiency Logic

Sources were included when they contributed directly to at least one of the following concerns: explainable NLP or faithful explanations; hallucination, factuality, or claim verification; retrieval, citation grounding, or long-context RAG; uncertainty or causal-language communication; human-AI evaluation, calibration, oversight, or responsible-AI documentation; or decision-support accountability. Foundational studies from before the LLM era were retained when they defined concepts still employed in the current evaluation (e.g., interpretability, rationales, factuality metrics, uncertainty communication, and automation accountability).

Domain-specific engineering and decision-support studies were not excluded solely because of their application domain. They were retained when NLP-assisted representations, textual evidence, explanation, grounding, uncertainty communication, or accountable decision support formed a substantive part of the method or deployment context. Studies reporting only predictive or optimization performance without a meaningful connection to language, evidence communication, or accountability were excluded. Empirical papers were also excluded if they only reported task performance without implications for explanation, grounding, factuality, citation support, human evaluation, or decision-facing communication.

The final reviewed set contains 104 sources, a scope chosen to balance breadth with analytical focus. It is sufficiently broad to cover the major strands required by the taxonomy while maintaining a clear focus on the intersection between explainable NLP and evidence-to-briefing accountability. The final source set was selected to provide coverage across the five taxonomy layers, established foundational work, recent 2025–2026 methodological developments, and representative engineering and regulatory applications. The set supports a targeted conceptual synthesis and should not be interpreted as an exhaustive bibliometric census of the entire field. After the final snowballing stage, newly inspected papers tended to reinforce existing categories rather than introduce additional layers beyond evidence representation, explanation generation, retrieval/source grounding, verification/evaluation, and human accountability.

Appendix A lists all 104 sources with their publication year, venue type, primary coding category, and role in the taxonomy (Table A1). Because several sources contributed to more than one stream, the coding was treated as a dominant-role classification for descriptive counting rather than as an exclusive characterization of each source (Table 3). For example, some RAG studies also contributed to factuality evaluation, and some responsible-AI or human-AI interaction studies also informed accountability. In such cases, each source was assigned to the primary category that best matched its central function within this review's taxonomy, and secondary relevance is discussed in the text as necessary.

2.4 Descriptive Source Statistics

Table 3 summarizes the primary coding distribution of the 104 reviewed sources. For descriptive counting, each source was assigned to one dominant category, although its secondary relevance to other research streams is discussed where appropriate.

Fig. 1 shows the annual distribution of the 104 reviewed sources across five major research streams. This distribution is descriptive, based on the primary coding reported in Appendix A, and should not be interpreted as a bibliometric estimate of the field as a whole.

Table 3: Primary coding distribution of the 104-source review set.

Primary Coding Category	No. of Sources
Explainable NLP, rationales, probing, and interpretability	19
Hallucination, factuality, and claim verification	15
RAG, citation grounding, retrieval adequacy, and long-context methods	26
Prompting, alignment, LLM evaluation, and preference optimization	22
Causal language, uncertainty communication, and accountable human decision support	22

Note: LLM: large language model, NLP: natural language processing, RAG: retrieval-augmented generation.

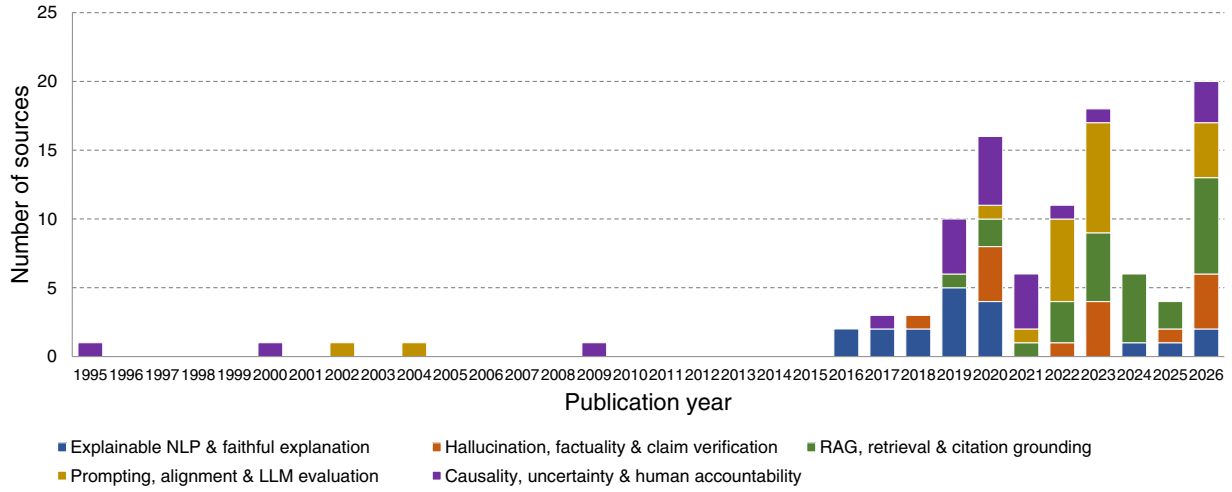

Figure 1: Annual publication volume across the five coded research streams in the reviewed source set ($n = 104$). Stacked bars show the number of sources assigned to each primary coding category by publication year. The distribution is descriptive and should not be interpreted as a measure of field-wide prevalence.

3 Task Formulation: Evidence-Grounded Decision Briefing

A clearer task formulation (Table 4) is necessary because the term "decision briefing" can become too broad. In this review, a briefing qualifies as an evidence-grounded decision briefing only when it meets four minimum conditions. First, the briefing is generated from identifiable evidence, including documents, tables, model outputs, retrieved passages, uncertainty intervals, attribution scores, rules, or contextual constraints. Second, the resulting text organizes these evidence elements into a reader-facing communication intended to support review or action. Third, the briefing contains at least one decision-relevant statement, such as a risk summary, recommended review priority, action boundary, uncertainty statement, or reason for escalation or hold. Fourth, each claim in the briefing must be traceable to supporting evidence, uncertainty information, applicable rules, and human review records.

This formulation distinguishes decision briefings from grounded summarization. Grounded summarization primarily assesses whether a summary is faithful to the source content. In contrast to grounded summarization, evaluating evidence-grounded decision briefings requires determining whether the generated text preserves uncertainty, uses causal wording appropriately, recommends only actions justified by evidence and rules, and assigns responsibility to the appropriate human or institutional actor. These briefings also differ from explanation generation: the briefing may include explanations, but its output is organized around decision-facing communication rather than explaining a model prediction.

Table 4: Task formulation for evidence-grounded decision briefings.

Component	Specification
Input	Type-labeled evidence, such as retrieved text, tables, model outputs, uncertainty estimates, attributions, rules, timestamps, user roles, and procedural context
Output	A natural-language briefing that summarizes the evidence, communicates uncertainty, explains key drivers, cites supporting sources, and may recommend a limited review action
Decision-facing element	At least one statement that may influence review, triage, escalation, hold, acceptance, or correction
Difference from grounded summarization	Extends beyond evidence summarization by incorporating action limits, causal restraint, uncertainty preservation, and responsibility assignment
Difference from explanation generation	May incorporate model explanations but evaluates the resulting text as a decision-facing and accountable artifact
Minimal accountability condition	Claims, citations, uncertainty, causal wording, recommended actions, and human review decisions must be traceable and open to inspection

Fig. 2 illustrates the task boundary of evidence-grounded decision briefings by distinguishing structured evidence inputs, the generated briefing artifact, and its decision-use context. This formulation differs from ordinary grounded summarization and model-explanation generation because it requires the resulting text to preserve source support, uncertainty, causal restraint, authorized action boundaries, and reviewability. Building on prior work in explainable NLP, faithful explanation, retrieval-augmented generation, and trustworthy retrieval, the proposed task narrows the focus to the joint evaluation of evidence, language, action, and human oversight in decision-facing communication [1,2,5,11].

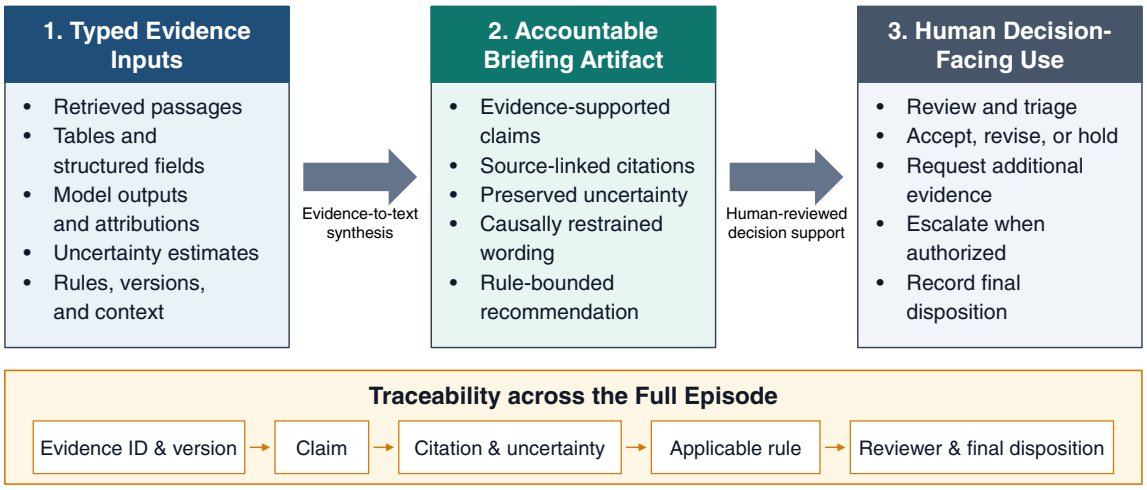

Figure 2: Task boundary of evidence-grounded decision briefings. The task begins with type-labeled evidence objects and produces a decision-facing textual artifact that supports review, triage, escalation, hold, acceptance, or revision.

4 Operational Taxonomy

The taxonomy integrates prior research on explainable NLP, hallucination and factuality, retrieval-augmented generation, and trustworthy retrieval into five operational layers [1–3,11]. These layers cover evidence representation, explanation generation, retrieval and source grounding, verification and evaluation, and human accountability. Together, they organize the literature according to how evidence is transformed into accountable decision-facing text.

Fig. 3 illustrates how the five layers progress from evidence representation to human review and accountability. Table 5 summarizes the representative methods, recurring failure modes, evaluation signals, and unresolved technical questions associated with each layer. The following section examines the supporting literature and explains why these layers should be integrated into a unified accountability framework for decision-facing briefings.

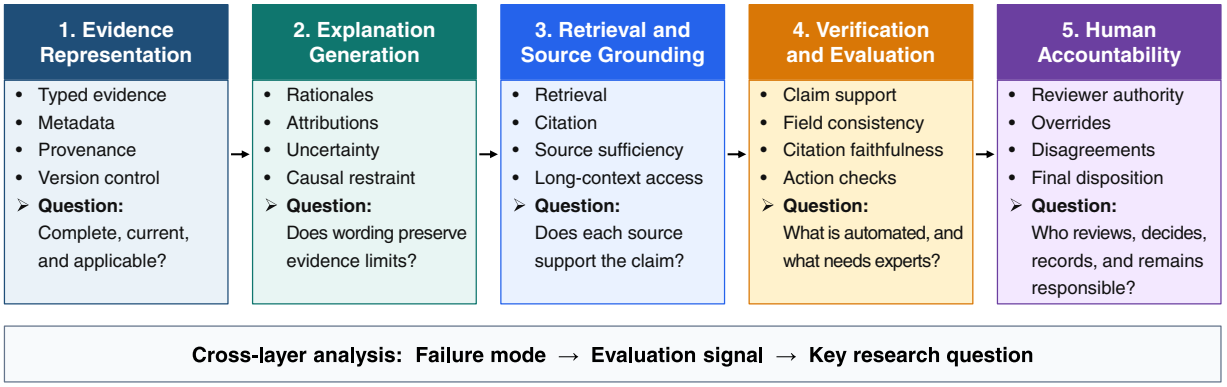

Figure 3: Five-layer taxonomy for explainable and accountable decision briefings, comprising evidence representation, explanation generation, retrieval and source grounding, verification and evaluation, and human accountability.

Table 5: Five-layer taxonomy with methods, failure modes, metrics, and open questions.

Layer	Representative Approaches	Main Failure Mode	Evaluation Signal	Key Research Question
Evidence representation	Type-labeled evidence, metadata, provenance tracking, and version control	Missing, outdated, mismatched, or inapplicable evidence	Evidence coverage, provenance completeness, and temporal consistency	How can evidence loss, version mismatch, and domain inapplicability be detected and prevented?
Explanation generation	Rationales, feature attributions, controlled generation, and intrinsic interpretability	Plausible but unsupported explanations or overstated causality	Explanation faithfulness and causal-language accuracy	How can explanations remain useful while preserving evidential support and epistemic restraint?
Retrieval and source grounding	RAG, citation generation, hierarchical retrieval, and long-context access	Relevant but insufficient, non-entailing, incomplete, or outdated sources	Retrieval utility, citation relevance, source sufficiency, and citation-to-claim entailment	How can source sufficiency, exception coverage, and temporal validity be evaluated reliably?

(Continued)

Table 5 (continued)

Layer	Representative Approaches	Main Failure Mode	Evaluation Signal	Key Research Question
Verification and evaluation	Claim segmentation, factuality checks, natural language inference pre-screening, structured-field validation, and expert review	Mixed claims, partial support, evaluator error, or unjustified action language	Claim support, field consistency, uncertainty preservation, and action appropriateness	Which evaluation dimensions can be automated, and which still require expert judgment?
Human accountability	Human-in-the-loop review, audit logs, role assignment, and override or disagreement records	Superficial review or unclear transfer of responsibility	Log completeness, justification of overrides, role clarity, and final disposition	How can human review produce meaningful and traceable institutional accountability?

Note: RAG: retrieval-augmented generation.

5 Critical Review of Major Research Streams

5.1 Explainability and Faithful Explanation

Explainable NLP provides a conceptual foundation for evidence-grounded briefings, but it does not fully resolve the language-layer problem: ensuring that generated explanations accurately represent the evidence and model behavior without overstating certainty, causality, or permissible action. Explanation faithfulness refers to the extent to which an explanation reflects the information or mechanisms that influenced a model's output rather than merely providing a plausible justification [5]. Rationale-based methods either select input passages as extractive rationales—text spans drawn directly from the input—or generate natural-language explanations intended to justify a prediction [12–15]. Faithfulness studies assess whether these explanations track the model's actual decision process instead of functioning as convincing post hoc accounts [16–18].

FaithLM is a model-agnostic framework for evaluating and improving the faithfulness of natural-language explanations generated by LLMs [19]. It treats faithfulness as an intervention-based property by testing whether contradicting an explanation changes the associated prediction. A recent survey situates this issue within the broader use of LLMs for explainable artificial intelligence and identifies continuing challenges related to faithfulness, user reliance, and accountability [20]. Collectively, these studies demonstrate that fluent and persuasive explanations cannot be assumed to reflect the evidence that influenced a model's output.

Decision briefings must also preserve the epistemic limits of the underlying evidence—that is, the boundaries of what the available evidence permits the system to claim. Post hoc attribution methods, such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), estimate the contributions of individual input features after a prediction has been generated [21,22]. Their outputs are attribution signals rather than direct evidence of causal mechanisms. Therefore, a high attribution score may justify the statement that a feature contributed to a prediction, but it cannot independently establish that the feature caused the observed event.

Broader interpretability research emphasizes rigorous evaluation, conceptual clarity, interpretable modeling in high-stakes settings, and human-centered explanation criteria [23–26]. Hase and Bansal [27] further show that explanations should be evaluated according to whether they help users anticipate model behavior and rely on it appropriately. Architecture-level transparency must also be distinguished

from explanation faithfulness: intrinsic interpretability embeds transparency within a model's structure or representations [8], whereas FaithLM evaluates whether generated explanations reflect information that influenced the prediction [19]. Consequently, an interpretable model does not necessarily produce a briefing that is evidence-faithful, causally restrained, or procedurally consistent with applicable rules and review responsibilities.

5.2 Hallucination, Factuality, and Source Fidelity

Hallucination refers here to generated content that is fabricated, internally inconsistent, or unsupported by the available evidence, whereas factuality concerns whether an individual claim is true or substantiated by a reliable source. Prior research has identified unsupported entities, fabricated relations, contradictions, and unwarranted specificity as recurring forms of hallucination [3,28–30]. Black-box hallucination detection evaluates these failures solely from model inputs and outputs, without access to internal parameters or activations [31]. This line of research establishes the need to evaluate generated text at the claim level rather than treat an entire briefing as uniformly reliable.

Truthfulness, atomic-fact, and large-scale hallucination benchmarks further demonstrate that fluent outputs may contain unsupported or unverifiable claims [32–34]. FActScore is a fine-grained factuality metric that decomposes long-form text into atomic facts and measures the proportion supported by a reliable knowledge source [33]. UNQOVER uses underspecified questions—questions lacking sufficient contextual information—to identify and quantify stereotyping biases and reasoning artifacts [35]. Together, these approaches show that factuality failures may result not only from fabricated content but also from unresolved ambiguity in the input or evaluation setting.

Recent studies published in 2026 further distinguish factual correctness from source fidelity, defined as the extent to which generated text accurately preserves the supplied evidence. Harmful Factuality Hallucination (HFH) describes cases in which a model "corrects" a perceived error in the source and produces an externally plausible but source-unfaithful output [36]. LLM-OASIS provides an end-to-end benchmark for passage-level factuality using controlled text manipulations and evidence-based verification [37]. Complementary research specifies design requirements for hallucination-detection benchmarks and introduces FactSearch, a reproducible agentic system in which autonomous components perform claim extraction, evidence retrieval, and claim-level verification [38,39].

Foundational claim-verification resources include Fact Extraction and VERification (FEVER), which labels general-domain claims using textual evidence, and SciFact, which pairs scientific claims with supporting or refuting abstracts and rationale passages [40,41]. These datasets provide structured settings for evaluating evidence retrieval, claim classification, and the textual rationales underlying each judgment. However, claim support alone is insufficient for decision-facing briefings. For example, evidence that a prediction interval has widened may justify reporting increased uncertainty, but it does not independently authorize immediate escalation. Accordingly, such briefings require separate assessments of rule compliance, uncertainty preservation, causal restraint, and action appropriateness.

5.3 RAG, Retrieval Adequacy, and Citation Faithfulness

RAG systems condition generation on evidence retrieved from external sources rather than relying solely on knowledge encoded in model parameters [2,42]. Hierarchical retrieval and long-context methods broaden access to information distributed across multiple documents, but access alone does not guarantee that the model will identify or use the decisive evidence [43,44]. Consequently, evaluation and robustness

studies examine retrieval relevance, effective context use, and resistance to irrelevant or distracting passages [45–48]. Recent reviews extend these concerns to trustworthy RAG, including reliability, hallucination mitigation, safety, explainability, and accountability across the retrieval–generation pipeline [11,49].

Foundational retrieval architectures include dense passage retrieval, retrieval-conditioned generation, retrieval-augmented language modeling, and demonstrate–search–predict approaches [50–53]. Attribution and citation-generation methods further link generated claims to identifiable sources [54–56]. Mallen et al. [57] distinguish parametric memory, encoded in model weights, from non-parametric memory supplied through retrieved evidence. Although these methods improve traceability, retrieval relevance does not establish source entailment, which requires the cited evidence to support the exact claim. Action authorization remains a separate question of whether the evidence and applicable rules permit the proposed response.

Recent LLM-oriented retrieval research further distinguishes human-perceived relevance from passage utility and distraction during downstream generation [58]. Utility and Distraction-aware Cumulative Gain (UDCG) evaluates retrieved passages according to both their usefulness and their potential to interfere with the downstream LLM. RAGVUE is a diagnostic, reference-free framework that separately assesses retrieval quality, answer relevance and completeness, claim-level faithfulness, and evaluator calibration [59]. This decomposition shows why a single aggregate score cannot identify whether a failure originates in retrieval, generation, grounding, or evaluation.

Citation faithfulness concerns whether a source supports the specific sentence or claim to which it is attached. This goes beyond citation presence because a topically relevant source may not entail the associated statement. CiteGuard evaluates citation attribution through retrieval-augmented validation rather than relying only on an unaided LLM judge [60]. VERICITE assesses sentence-level citation support in retrieval-augmented medical answers [61]. Mr Dre, an evaluation suite for Deep Research Agents (DRAs), extends this analysis to multi-turn report revision and shows that iterative editing can weaken content coverage, claim–citation alignment, and citation quality [62].

5.4 Causal Language and Uncertainty Communication

Causal-language discipline requires generated wording to reflect the type and strength of the available evidence. Causal inference examines whether changing one factor would alter another under explicit assumptions, whereas causal NLP applies causal principles to textual data and language-model behavior [63–66]. Causal mediation analysis further investigates whether an observed effect operates through identifiable intermediate components or representations within a model [67]. These distinctions are essential for decision briefings because predictive association, feature attribution, and causal effect provide different levels of evidential support.

Generated briefings may weaken or omit epistemic qualifiers—words or phrases that indicate uncertainty, limitations, or evidential status. For example, “the model attributed the prediction to X” may be rewritten as “X caused the event,” while confidence intervals, data-quality warnings, or alternative explanations may be removed. Accordingly, evaluation should include controlled contrast pairs that differ primarily in causal or epistemic strength, such as “caused” vs. “was associated with” and “contributed to” vs. “was consistent with.” These tests assess whether generated wording remains calibrated to the evidence rather than becoming more certain or causal than the evidence permits.

Uncertainty communication concerns how probabilities, intervals, limitations, and confidence are expressed so that readers neither overlook risk nor infer unwarranted certainty [68,69]. Recent methods extend uncertainty quantification beyond a single self-reported confidence score. Evidential Semantic

Entropy (EVSE) captures uncertainty from unobserved answer possibilities and semantic relations among observed responses [70], while Faithfulness-aware Retrieval-Augmented UNcertainty Quantification (FRANQ) distinguishes factual correctness from faithfulness to retrieved evidence in RAG outputs [71]. Therefore, decision-briefing evaluation should assess whether uncertainty is preserved, omitted, exaggerated, understated, or misleadingly expressed.

5.5 Human Evaluation and Accountability

Human evaluation and preference optimization use comparative judgments to guide model behavior, but these signals are neither neutral nor sufficient for full accountability. Reinforcement learning from human feedback (RLHF) trains reward models from human preferences, whereas Constitutional AI relies on written principles and AI-generated critiques or preferences to shape safer behavior [72,73]. Helpful–harmless assistant training and WebGPT extend these approaches to safety and source-supported question answering, yet outputs remain sensitive to prompts, demonstrations, evaluator preferences, and other inference-time context [74–76].

General model reports and broad challenge surveys document model capabilities, safety constraints, factual-grounding limitations, and deployment risks [77–80]. Evidence from medical question answering further shows that strong performance in a high-stakes domain still requires domain-specific validation and careful interpretation [81]. These studies establish important alignment and evaluation baselines, but they do not determine whether each briefing claim is supported by its cited evidence. Nor can they independently assess whether uncertainty, causal wording, and proposed actions are appropriate to the available evidence and decision context.

The term LLM-as-a-judge refers to the use of a language model to rate or compare outputs according to specified criteria. MT-Bench evaluates multi-turn responses, whereas Chatbot Arena uses crowdsourced pairwise comparisons; the original study also identifies position, verbosity, and self-enhancement biases [82]. G-Eval assesses natural language generation outputs using explicit criteria [83], while GPTScore uses natural-language instructions to define the quality dimensions being evaluated [84]. Although these approaches improve evaluation scalability, their scores should remain supplementary to source-entailment checks, uncertainty preservation, causal restraint, action safety, and expert review.

Generated reasoning traces should not be treated as accountability evidence by default. Chain-of-thought (CoT) prompting elicits intermediate reasoning steps, while self-consistency samples multiple reasoning paths and selects a convergent answer [85,86]. These traces may still reflect post hoc rationalization, shared assumptions, or prompt-induced artifacts. Therefore, greater detail or agreement across reasoning paths does not establish explanation faithfulness, evidential sufficiency, causal validity, or action authorization. Moreover, systematic prompting research shows that outcomes vary with task formulation, example selection, and decoding strategy [87].

Documentation and governance research provides complementary accountability foundations through Model Cards, Datasheets for Datasets, and analyses of systemic risks associated with large language models [88–90]. Human-centered studies show that explanation use varies with task design, practitioner understanding, and stakeholder role [91–93]. Human–AI interaction guidelines, cognitive-forcing interventions, levels-of-automation research, and situation-awareness theory further indicate that meaningful oversight depends on interface design, reviewer authority, and decision context [94–97]. Therefore, accountability should be implemented as a traceable review process that records evidence, generated claims, citations, uncertainty statements, reviewer identities, overrides, disagreements, and final decisions, as summarized in Table 6.

Table 6: Mapping of reviewed research streams to the five-layer taxonomy.

Research Stream	Main Taxonomy Layers	Primary Contribution	Remaining Gap for Decision Briefings
Explainable NLP and faithful explanation [1,5,19,27]	Explanation generation; verification and evaluation	Defines explanation faithfulness, rationale quality, model–behavior alignment, and user-oriented usefulness	Does not by itself establish exact source support, permissible action language, or institutional responsibility
Hallucination, factuality, and claim verification [3,33,36,37]	Verification and evaluation	Provides hallucination taxonomies, atomic-claim evaluation, source-fidelity distinctions, and factuality benchmarks	Does not fully address causal restraint, uncertainty communication, or action authorization
RAG, retrieval, and citation grounding [2,11,58,61]	Evidence representation; retrieval and source grounding; verification and evaluation	Improves evidence access and evaluates retrieval utility, citation relevance, and claim-level support	Retrieval quality and citation presence do not guarantee exact entailment, exception coverage, temporal validity, or action authorization
Causal language and uncertainty [63,68,70,71]	Explanation generation; verification and evaluation	Clarifies causal standards, epistemic limits, risk framing, and uncertainty communication	Often remains weakly connected to retrieval quality, action rules, and auditability
Human–AI accountability and oversight [88,93,95,97]	Human accountability	Provides principles for documentation, stakeholder roles, overreliance, reviewer authority, and situational awareness	Requires stronger links to claim-level outputs, evaluator reliability, overrides, and final decision records
Engineering and regulatory decision support [98–100]	Evidence representation; retrieval and source grounding; verification and evaluation; human accountability	Demonstrates how predictive outputs and fragmented regulatory evidence can inform operational judgments	Predictive accuracy or relevant retrieval alone does not establish applicability, exception coverage, or action authorization

Note: AI: artificial intelligence, NLP: natural language processing, RAG: retrieval-augmented generation.

5.6 Engineering and Regulatory Decision Support

Accountable NLP is also relevant in engineering settings, where model outputs may influence inspection, maintenance, or operational restrictions. Engineering generalizability refers here to whether the proposed accountability framework remains applicable when evidence types, prediction targets, and decision

rules differ from those used in conventional NLP benchmarks. Wang et al. [98] applied NLP-assisted transfer learning to predict creep strain across diverse adhesively bonded joints. However, their study is treated as an application case rather than a direct validation of the framework because its primary contribution lies in predictive modeling rather than accountability evaluation.

The source domain is the setting used for initial learning, whereas the target domain is the new setting to which the learned representation is adapted. Accordingly, a decision briefing should report the joint configuration, source- and target-domain provenance, limitations of the target-domain data, uncertainty, inspection history, and the rule governing maintenance or operational restrictions. Predictive accuracy alone cannot establish that a component will fail or that a specific intervention is authorized. Instead, it may justify a bounded review trigger when a predefined engineering threshold is met.

Regulatory decision support introduces related but distinct requirements. Statute-centric question answering (QA) addresses questions whose controlling evidence lies primarily in statutes or regulations rather than case law. SearchFireSafety evaluates whether systems can retrieve hierarchically fragmented fire-safety provisions and safely abstain when the statutory context is incomplete [99]. Similarly, a recent survey of legal LLM agents identifies hallucination, outdated information, verifiability, tool use, and agent-specific evaluation as major deployment challenges [100].

Taken together, these cases show that evidence-grounded accountability extends beyond generic risk scores. Relevant evidence may be technically accurate yet incomplete, outdated, inapplicable to a specific case, or insufficient to authorize an action. Therefore, the proposed framework evaluates predictive or retrieval performance separately from applicability, exception coverage, uncertainty communication, and human authorization. Finally, Table 7 complements this cross-domain synthesis by presenting selected quantitative evidence on current limitations in factuality, retrieval, and citation evaluation.

Table 7: Targeted quantitative evidence from recent factuality, retrieval, and citation-evaluation studies.

Evaluation Concern	Study	Reported Result	Interpretation
Source fidelity	Harmful Factuality [36]	A simple instructional prompt reduced Harmful Factuality Hallucination rates by approximately 50%.	Factual correctness and fidelity to the supplied source are distinct and separately measurable.
End-to-end factuality	LLM-OASIS [37]	GPT-4o achieved up to 60% accuracy on the proposed end-to-end factuality-evaluation task.	Strong contemporary models still face substantial difficulty in evaluating the factuality of complete passages.
Retrieval evaluation	UDCG [58]	Correlation with end-to-end RAG answer accuracy improved by up to 36% over traditional retrieval metrics.	Human-oriented relevance metrics may not reflect how an LLM uses helpful and distracting passages.

(Continued)

Table 7 (continued)

Evaluation Concern	Study	Reported Result	Interpretation
Citation-judge reliability	CiteGuard [60]	LLM-based citation attribution produced recall as low as 16%–17% in the reported evaluation settings.	Automated citation judgments require retrieval support, calibration, and human validation.
Sentence-level citation support	VERICITE [61]	Only approximately 27%–41% of citation–sentence pairs were supported at retrieval depths of 3 and 5.	Retrieving relevant biomedical documents does not guarantee support for each associated sentence.

Across these heterogeneous tasks, a consistent pattern emerges: strong generators and automated evaluators continue to exhibit weaknesses in source fidelity, factuality, retrieval alignment, and citation support. Accordingly, deterministic checks can verify directly testable properties, such as identifiers, dates, units, and source versions, through predefined rules. Automated semantic pre-screening can then use learned models to flag potential entailment, contradiction, or grounding failures before human review. However, expert review remains necessary when evidential support is partial, relevant information is distributed across multiple sources, or an action depends on domain-specific rules and professional judgment.

6 Evidence Grounding and Operational Evaluation

Evidence grounding requires more than retrieving material that is topically related to a request. Recent research on trustworthy RAG, retrieval evaluation, and citation verification shows that topical relevance alone cannot establish exact source support or action permissibility [11,49,58,61]. Specifically, retrieval adequacy asks whether the evidence needed to assess a claim has been retrieved, whereas source entailment asks whether that evidence supports the claim's exact wording. Beyond these requirements, action authorization examines whether the supported claim, together with the applicable rule, permits the proposed review priority or action.

6.1 Retrieval Adequacy, Source Entailment, and Action Authorization

A recurring problem in accountable generation is the conflation of retrieval with grounding. Recent reviews of trustworthy RAG and hallucination show that retrieving topically relevant material does not by itself ensure reliable or adequately supported generation [11,49]. Moreover, LLM-oriented retrieval and citation studies distinguish the downstream utility of retrieved passages from exact sentence-level support [58,61]. Accordingly, this review organizes grounding into six progressively stricter evidential checks that should be evaluated separately rather than collapsed into a single score, as summarized in Table 8.

Retrieval recall asks whether the evidence needed to answer a request has been retrieved. Next, citation relevance checks whether the cited source is related to the claim, whereas citation-to-claim entailment determines whether it directly supports the claim's exact wording. Temporal validity and exception coverage then assess whether the current source version, relevant limitations, decision thresholds, and counterconditions have been identified. Finally, action authorization applies the strictest test by determining whether the supported claim and governing rule actually permit the proposed priority or response.

Table 8: Distinguishing retrieval adequacy from source entailment and action authorization.

Level	Question	Possible Evaluation Signal	Importance
Retrieval recall	Was the material needed to address the request retrieved?	Top-k recall and answer-containing passage recall	Access is necessary but does not show that a cited passage supports a generated claim.
Citation relevance	Is the cited source topically related to the associated claim?	Citation relevance and source-topic matching	Topical relevance does not imply exact support or procedural applicability.
Citation-to-claim entailment	Does the cited source support the exact generated claim?	Entailed, contradicted, incomplete, or irrelevant labels	Requires claim segmentation and comparison with the cited evidence.
Temporal validity	Is the source valid for the relevant date, version, jurisdiction, policy, or data release?	Effective-date checks, version metadata, and source freshness	A historically true statement from a superseded source can still be unsafe.
Exception coverage	Were relevant exceptions, thresholds, definitions, footnotes, and counterconditions retrieved and used?	Exception-recall and rule-condition-completeness labels	Missing exceptions can reverse the permissible conclusion or action.
Action authorization	Do the available evidence and the applicable rule jointly justify the proposed review priority or action?	Authorized, unauthorized, over-escalated, or insufficient-evidence labels	Action permissibility cannot be inferred from factuality or citation relevance alone.

Note: Top-k recall measures whether the required evidence appears among the top k retrieved items.

Fig. 4 illustrates the cumulative progression from retrieval recall to action authorization. At each successive tier, additional requirements are introduced for exact claim support, temporal validity, exception coverage, and procedural permissibility. Therefore, a source may be relevant and factually supportive yet still be insufficient to authorize the proposed action.

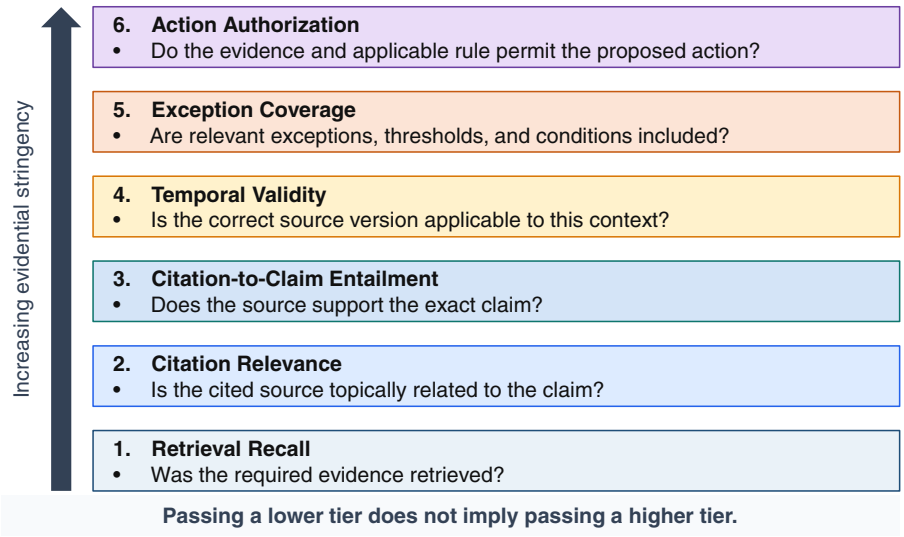

Figure 4: Conceptual hierarchy from retrieval recall to action authorization. Each successive tier imposes a more stringent evidential requirement, progressing from access to relevant material to claim support, temporal applicability, exception coverage, and authorization of the proposed action.

6.2 Operational Evaluation Framework

Fig. 5 translates the taxonomy into a high-level structure for future benchmark design, while Table 9 specifies the corresponding evaluation units, labels, automated checks, and expert-review triggers. This framework supports the systematic assessment of evidence-grounded decision briefings across different domains and decision contexts. Because it remains a proposed operational structure rather than a validated protocol, the framework requires domain-specific annotation guidelines, inter-rater reliability testing, and empirical validation. Its purpose is to distinguish properties that can be verified deterministically from those requiring semantic interpretation, contextual assessment, or professional judgment. The five central evaluation dimensions are operationalized through the corresponding components in Table 9, with claim support and field consistency included as prerequisite checks for reliable evaluation.

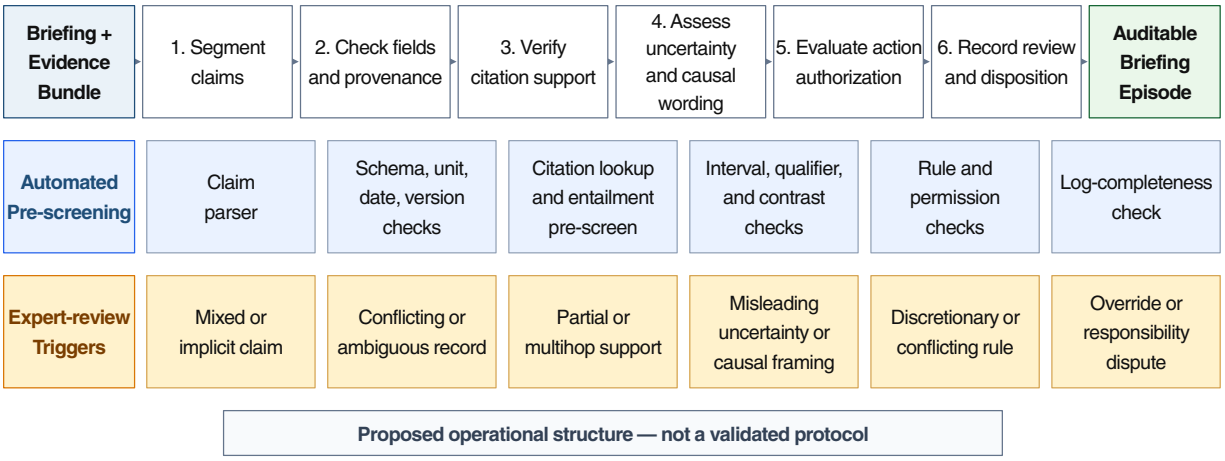

Figure 5: Simplified evaluation workflow for accountable decision briefings. The workflow reduces a generated briefing into inspectable claims and evaluates evidence support, language discipline, action appropriateness, and review accountability.

Table 9: Operational evaluation components for evidence-grounded decision briefings.

Dimension	Unit	Labels	Automatic Checks	Expert Triggers
Claim support	Atomic claim	Supported, contradicted, incomplete, or unverifiable	Citation lookup and NLI pre-screen	Implicit, multihop, or mixed claim
Field consistency	Number, unit, date, identifier, and source version	Match, mismatch, stale, or missing	Regex, schema, unit, and version checks	Ambiguous rounding or conflicting records
Citation faithfulness	Citation–claim or citation–sentence pair	Entailing, partially supporting, relevant only, fabricated, or inaccessible	Citation existence and source retrieval	Table-heavy, legal, scientific, or multihop source
Uncertainty preservation	Interval, probability, confidence category, limitation, or qualifier	Preserved, omitted, exaggerated, understated, or misframed	Interval, threshold, and qualifier checks	Misleading verbal framing or user-interpretation risk
Causal restraint	Causal or attributional phrase	Causal, associative, attributional, appropriately hedged, or overclaimed	Verb, modality, and contrast-pair checks	Causal-evidence and domain judgment
Action appropriateness	Action label, recommendation, or escalation priority	Authorized, unsupported, over-escalated, under-escalated, or insufficient evidence	Rule metadata and permission checks	Rule–evidence conflict or discretionary professional judgment
Human accountability	Briefing episode	Complete, partial, disputed, or unclear	Log, identity, role, and disposition completeness	Responsibility dispute, override disagreement, or institutional-policy conflict

Note: NLI: natural language inference, used here as an automated pre-screen rather than a definitive entailment judgment.

Common baseline metrics include Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), BERTScore, and Sentence-BERT [101–104]. BLEU and ROUGE primarily measure lexical overlap, whereas BERTScore and Sentence-BERT use contextual representations to estimate semantic similarity. Although useful for comparing surface form and semantic content, these metrics do not directly assess source entailment, temporal validity, exception coverage, or action authorization. Accordingly, the proposed framework treats claims, citations, uncertainty statements, causal wording, and proposed actions as distinct evaluation units.

Evaluation begins by segmenting a briefing into atomic or otherwise reviewable claims so that each statement can be assessed against its supporting evidence. Next, the framework examines source support, field consistency, citation faithfulness, uncertainty preservation, causal restraint, and action appropriateness. It then evaluates whether the human-review record adequately documents reviewer identity, interventions,

disagreements, overrides, and final decisions. Together, these stages clarify how deterministic checks, semantic evaluation, and expert review can be combined without reducing accountability to a single aggregate score.

7 Illustrative Use Cases

The three cases in Table 10 illustrate how the proposed framework can be applied across generic, engineering, and regulatory settings. They are pedagogical examples intended to clarify the framework's operation and should not be interpreted as empirical validation.

Table 10: Illustrative applications of the accountability framework.

Use Case	Step	Illustrative Content	Review Interpretation
Generic risk scoring	Evidence	Risk score = 0.72; prior baseline = 0.41; interval = 0.60–0.81; top attributions = missing sensor data and recent demand change; escalation rule requires verified data freshness; freshness warning is active.	Combines numerical evidence, uncertainty, attribution signals, rule conditions, and a data-quality exception.
	Generated briefing	Current risk is elevated (0.72 vs. a baseline of 0.41; interval 0.60–0.81). The model attributes the increase to missing sensor data and recent demand changes. Because data freshness is not verified, hold the case for human review before escalation.	Preserves the interval, uses attributional rather than causal language, and bounds the action by the applicable exception.
	Action judgment	Hold for human review before escalation.	Authorized because the rule requires freshness verification; a high score alone would not authorize escalation.
Engineering maintenance [98]	Evidence	Predicted creep-strain trajectory, uncertainty, joint and material configuration, source- and target-domain data provenance, target-domain sample limitations, inspection history, and maintenance threshold.	Makes applicability and engineering-rule information explicit rather than treating the prediction as a self-authorizing action.

(Continued)

Table 10 (continued)

Use Case	Step	Illustrative Content	Review Interpretation
	Accountable briefing	The model indicates elevated creep-strain risk for this joint configuration. Because the target-domain evidence is limited, treat the result as an inspection trigger rather than proof of imminent failure. Apply an operational restriction only if the governing engineering threshold is met.	Separates prediction from observed failure, preserves domain limitations, and conditions the action on the governing rule.
Regulatory fire safety [99,100]	Evidence	Retrieved statutory provisions, definitions, hierarchy links, effective dates, exceptions, thresholds, jurisdiction, and retrieval completeness flags.	Regulatory evidence can be fragmented across hierarchically linked provisions; topical retrieval alone is insufficient.
	Accountable briefing	The retrieved provisions do not include the exception governing this configuration. No approval, rejection, or enforcement recommendation should be issued until the missing statutory context is retrieved and reviewed.	Uses safe abstention when the evidence is incomplete and avoids converting partial legal retrieval into an unauthorized action.

7.1 Generic Risk-Scoring Example

A decision briefing should not recommend approval, rejection, enforcement, or operational action unless the controlling exception, decision threshold, and valid source version have been established. Similarly, routine processing should not proceed when a current document is required but only an outdated record is available, because topical relevance does not establish procedural validity. For this reason, the briefing should identify the missing or outdated evidence, explain why authorization cannot yet be determined, and route the case for human review.

7.2 Engineering Maintenance Use Case

In the engineering case, an NLP-assisted transfer-learning system predicts elevated creep strain in an adhesively bonded joint [98]. The briefing should report the joint and material configuration, source- and target-domain data, predicted strain, uncertainty, applicability limits, inspection history, and governing maintenance threshold. Under these conditions, the prediction should be presented as a trigger for inspection rather than proof of imminent failure, and an operational restriction should be recommended only when the applicable threshold and supporting evidence authorize that action.

7.3 Regulatory Fire-Safety Use Case

In the regulatory case, fire-safety requirements may be distributed across hierarchically linked provisions, definitions, thresholds, and exceptions. SearchFireSafety evaluates whether systems can retrieve this fragmented statutory evidence and abstain safely when the available context is incomplete [99]. In regulatory settings, legal-agent systems should avoid unsupported approval or enforcement recommendations and preserve source versions, retrieved provisions, tool actions, reviewer interventions, and final dispositions to support verification and institutional accountability [100].

8 Human-Accountability Structures

Human accountability requires more than a generic human-in-the-loop checkpoint. Documentation frameworks emphasize the need to record model purpose, dataset provenance, intended use, limitations, and foreseeable risks [88–90]. Human-centered studies further show that explanations and interpretability tools are understood and applied differently depending on task design, practitioner expertise, and stakeholder role [91–93]. Research on human–AI interaction, cognitive forcing, levels of automation, and situation awareness also indicates that meaningful oversight depends on interface design, reviewer authority, and the surrounding decision context [94–97].

An accountability record should capture four connected categories of information. Evidence records should include source identifiers, versions, timestamps, retrieval outputs, uncertainty fields, and applicable rules. System records should preserve the generated briefing, segmented claims, automated checks, citation results, exception flags, and proposed action labels. Human-review records should document reviewer identity, authority, interventions, disagreements, overrides, and final decisions. Downstream records should capture the implemented action and subsequent outcomes, when available. Fig. 6 summarizes the minimum audit trail and the relationships among these records.

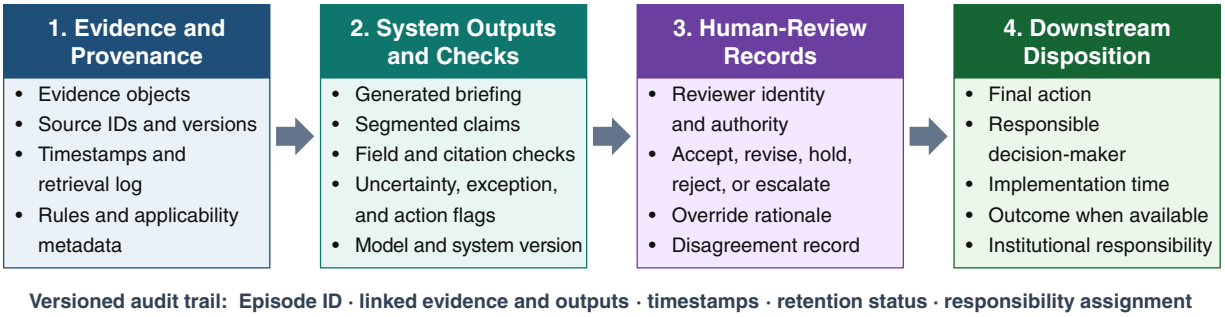

Figure 6: Minimum audit trail for human-reviewed decision briefings, identifying records that should be retained for ex post review: evidence, retrieval log, generated briefings, claim-level checks, reviewer decisions, override or disagreement reasons, final actions, and institutional responsibility.

Responsibility should be assigned explicitly rather than attributed collectively to an undifferentiated human–AI team. The system should preserve its inputs, outputs, retrieval history, model version, and automated-check results. The reviewer should evaluate the evidence and document the rationale for accepting, revising, holding, or rejecting the proposed action. The institution should define reviewer authority, escalation procedures, permissible actions, record-retention periods, audit obligations, and responsibility for downstream consequences, as summarized in Table 11.

Table 11: Concrete accountability structures for decision-facing natural language processing briefings.

Accountability Element	Minimum Record	Purpose
Evidence logging	Type-labeled evidence, source IDs, retrieval outputs, timestamps, rule versions, uncertainty fields, and applicability metadata	Enables reconstruction of the information available to the system and reviewer
System-output logging	Generated briefing, segmented claims, field-check results, citation checks, uncertainty and exception flags, and recommended action labels	Makes automated processing and failure points inspectable
Reviewer role	Reviewer identity, professional role, authority level, and relevant expertise	Clarifies who was qualified and authorized to exercise judgment
Override recording	Reason for accepting, changing, rejecting, holding, or escalating contrary to the system recommendation	Prevents silent disagreement and supports later audit
Disagreement handling	Whether disagreement concerns evidence, wording, citation support, uncertainty, causal language, action label, or rule interpretation	Separates failures in generated language from domain-rule disputes
Downstream disposition	Final action, responsible decision-maker, implementation time, and outcome when available	Connects the briefing and review episode to the real-world decision
Audit trail	Versioned evidence, generated text, claim-level check results, reviewer decisions, overrides, disagreements, final actions, and model or system versions	Supports ex post error analysis, reproducibility, and responsibility tracing
Responsibility assignment	Defined roles for the system, reviewer, institution, and downstream decision-maker	Prevents nominal human approval from obscuring actual decision authority and accountability

9 Open Technical Challenges

9.1 Claim Segmentation and Claim Typing

Generated briefings often combine factual, numerical, attributional, causal, uncertainty-related, procedural, and action-oriented claims within a single sentence. FActScore demonstrates the value of decomposing

long-form text into atomic facts, while FEVER and SciFact provide structured evidence labels and supporting rationales for general-domain and scientific claims [33,40,41]. Future benchmarks should extend these approaches by assigning a claim type before assessing evidence support, uncertainty preservation, causal restraint, or action appropriateness. Inter-annotator agreement—the consistency of labels assigned by different annotators—should be reported separately for each claim type because segmentation or typing errors can propagate through all subsequent evaluation stages.

9.2 Semantic Verification beyond Field Checks

Deterministic checks are appropriate for numbers, dates, units, identifiers, source versions, and controlled action labels that can be directly matched against structured evidence. Natural language inference (NLI), citation-validation systems, and expert review are required when evaluation involves paraphrases, implied conclusions, source-to-claim entailment, or multihop evidence [60,61]. MT-Bench and Chatbot Arena approximate human preferences at scale but their results remain vulnerable to position, verbosity, self-enhancement, and reasoning-related biases [82], while G-Eval [83] and GPTScore [84] provide scalable evaluations of generation quality rather than definitive accountability judgments. Accordingly, future protocols should separate rule-verifiable properties from semantic judgments and explicitly record uncertainty, evaluator disagreement, and cases requiring professional review.

9.3 Causal-Language Test Suites

Causal-language test suites should include controlled contrast pairs that differ primarily in evidential strength, such as caused vs. was associated with and contributed to vs. was consistent with. Causal inference and causal NLP show that causal statements require assumptions and evidence beyond predictive association, feature attribution, or textual correlation [63–66]. Causal mediation analysis provides an additional basis for testing whether evidence about intermediate model components is represented accurately rather than converted into unsupported causal claims [67]. Evaluations should measure both causal overstatement and excessive hedging because either failure can distort the interpretation of evidence and lead to inappropriate decision-facing communication.

9.4 Long-Context RAG and Exception Coverage

Hierarchical retrieval and long-context models can improve access to evidence distributed across multiple passages, documents, or linked provisions [43,44]. Nevertheless, utility-aware retrieval and sentence-level citation evaluation show that broader access does not guarantee effective use, exact support, or identification of controlling evidence [58,61]. Consequently, future benchmarks should include exceptions, footnotes, temporal cutoffs, superseded versions, and decision thresholds whose omission changes the permissible response. Statute-centric settings are particularly useful because SearchFireSafety directly tests hierarchically fragmented retrieval and safe abstention when controlling provisions, definitions, or exceptions are missing [99].

9.5 Uncertainty Preservation and Calibrated Trust

Decision briefings should communicate probabilities, intervals, limitations, and data-quality warnings without obscuring material risk or creating unwarranted alarm [68,69]. Evidential Semantic Entropy accounts for uncertainty arising from unobserved answer possibilities and semantic relations among observed responses, while faithfulness-aware uncertainty quantification distinguishes general factual confidence from support grounded in retrieved evidence [70,71]. User studies should separately assess readability, comprehension, uncertainty calibration, and appropriate reliance rather than treating perceived clarity as

evidence of trustworthy communication. Interface mechanisms should also be evaluated because cognitive-forcing interventions that require active user reasoning may reduce overreliance more effectively than passive explanations alone [95].

9.6 Human Evaluation without Preference Bias

Preference-alignment methods can improve instruction following, helpfulness, harmlessness, and source-supported response generation [72–75]. However, success on these objectives does not by itself guarantee that each claim is supported by its cited source, that uncertainty and causal boundaries are preserved, or that a proposed action is safe and consistent with domain-specific rules. Accordingly, LLM-based evaluators such as MT-Bench [82], G-Eval [83], and GPTScore [84] should be reported separately from evidence-grounding and action-authorization assessments, because general quality scores may obscure failures in support or procedural appropriateness. Human evaluation should also measure reviewer disagreement, overreliance, sensitivity to fluent wording, and willingness to challenge recommendations that appear persuasive but remain insufficiently supported [95].

9.7 Multilingual and Domain-Specific Accountability

Multilingual LLM research identifies continuing challenges in cross-lingual knowledge transfer, language coverage, fairness, and accessibility [10]. Decision-briefing benchmarks should test whether source support, uncertainty, causal restraint, and action boundaries remain intact during translation, multilingual retrieval, or the integration of evidence written in different languages. Domain-specific validation should also account for jurisdictional rules, technical terminology, formulas, tables, and professional review standards, as illustrated by statute-centric legal QA and legal-agent applications [99,100]. Cross-domain evaluation should preserve common accountability requirements while allowing evidentiary thresholds, permissible actions, and reviewer responsibilities to vary across languages, jurisdictions, and institutional settings.

10 Conclusions, Limitations, and Future Research Agenda

10.1 Main Conclusions

This review positions evidence-grounded decision briefings as a distinct problem in accountable NLP. The synthesis demonstrates that fluent generation, factual plausibility, successful retrieval, and citation presence are each insufficient to ensure decision-facing reliability. Instead, an accountable briefing must preserve the relationships among type-labeled evidence, atomic claims, exact source support, uncertainty, causal wording, permissible actions, and identifiable human responsibility. The quantitative comparisons further indicate that even strong generators and automated evaluators continue to exhibit measurable weaknesses in source fidelity, factuality, retrieval alignment, and citation support.

The proposed five-layer taxonomy integrates research streams that have largely developed in separate communities. Explainable NLP contributes methods for rationale generation and explanation faithfulness; hallucination and factuality research provides claim-level verification tools; RAG and citation research supports external evidence access and attribution; causal and uncertainty research defines epistemic boundaries; and human–AI research informs oversight and responsibility. The engineering and regulatory cases further show that these requirements remain relevant when evidence types, applicability conditions, and decision rules differ from those used in general NLP benchmarks. Therefore, the framework's primary contribution lies in organizing these elements into a unified evidence-to-action accountability process.

10.2 Limitations and Practical Implications

This article presents a targeted narrative synthesis rather than a systematic review or meta-analysis. Although the 104-source set covers a broad range of relevant research, it is not exhaustive, and primary-category coding cannot capture every secondary contribution. The quantitative comparisons draw on heterogeneous datasets, models, metrics, and annotation procedures and should not be interpreted as pooled effects or direct system rankings. The proposed taxonomy, labels, use cases, and evaluation structure also require empirical validation before they can be adopted as standardized evaluation tools.

Nevertheless, the framework offers practical value across three stages of system development and use. At design time, developers can specify the required evidence, provenance, uncertainty, rule, and reviewer-role fields. At runtime, the framework separates deterministic checks from semantic pre-screening and expert judgment. At the governance stage, it identifies the records needed to reconstruct a briefing episode, including retrieved sources, generated claims, automated checks, reviewer decisions, overrides, disagreements, and final actions.

However, the taxonomy should not be implemented as a universal scoring template without domain adaptation. Engineering, healthcare, legal, compliance, financial, and public-sector systems differ in their evidentiary standards, permissible actions, reviewer qualifications, escalation thresholds, and retention obligations. Accordingly, deployment requires domain-specific annotation guidelines, action rules, expert-review criteria, reliability testing, and validation with representative samples of users, evidence types, and decision cases.

10.3 Priority Future Research Directions

Future research should prioritize four connected directions. First, benchmark datasets should jointly annotate claims, citations, uncertainty statements, causal wording, action labels, and reviewer records while reporting annotation reliability separately for each claim type. Second, retrieval evaluation should move beyond topical relevance to assess exact citation-to-claim entailment, temporal validity, exception coverage, and action authorization. Third, human evaluation should distinguish preferences for fluent or confident text from appropriate reliance, causal restraint, disagreement handling, and action safety. Fourth, the framework should be externally validated in multilingual and domain-specific settings through multi-institutional studies that examine both technical performance and audit-record completeness. Together, these priorities would move accountable NLP beyond loosely connected metrics toward a testable and auditable discipline for evidence-grounded decision communication.

Acknowledgement: The author sincerely thanks Prof. Dr. Jehyeok Rew of Duksung Women's University for his insights into natural language processing and large language models, and Prof. Dr. Seungmin Rho of Chung-Ang University for his perspectives on trustworthy and explainable artificial intelligence.

Funding Statement: The authors received no specific funding for this study.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: Given his role as an Editorial Board Member of this journal, Jihoon Moon had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The author declares no other conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
LLM	Large language model
NLP	Natural language processing
RAG	Retrieval-augmented generation
RLHF	Reinforcement learning from human feedback

Appendix A Source-Level Coding Table

Appendix A provides source-level coding for the 104 studies included in this narrative review. Source numbers correspond to the final reference list in order of first appearance. Venue type refers to the specific version cited, and parallel preprint, conference, or journal versions are not counted more than once. The coding distinguishes journal articles, conference papers, preprints, technical reports, and books. The "Role in taxonomy" column summarizes each source's conceptual contribution and should not be interpreted as evidence of a pooled quantitative effect.

Table A1: Source-level coding of the 104 sources included in the narrative review.

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
1	Danilevsky, Qian, Aharonov, Katsis, Kawas, Sen	2020	Conference paper	Explainable NLP, rationales, probing, and interpretability	Differentiation from prior explainable NLP surveys
2	Gao, Xiong, Gao, Jia, Pan, Bi, et al.	2023	Preprint	RAG, citation grounding, retrieval adequacy, and long-context methods	Differentiation from general RAG surveys
3	Ji, Lee, Frieske, Yu, Su, Xu, et al.	2023	Journal article	Hallucination, factuality, and claim verification	Differentiation from hallucination surveys
4	Kotonya, Toni	2020	Conference paper	Hallucination, factuality, and claim verification	Explainable fact-checking and claim verification
5	Lyu, Apidianaki, Callison-Burch	2024	Journal article	Explainable NLP, rationales, probing, and interpretability	Differentiation from faithful-explanation surveys
6	Qin, Chen, Feng, Wu, Zhang, Li, et al.	2026	Journal review article	Prompting, alignment, LLM evaluation, and preference optimization	Updated survey of LLM applications and adaptation paradigms across NLP
7	Javed, Shah, Ali, Kwak	2026	Journal review article	Prompting, alignment, LLM evaluation, and preference optimization	Recent LLM evolution, architectural trends, deployment challenges, and accountability context

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
8	Gao, Meng, Zhou, Pan	2026	Conference survey paper	Explainable NLP, rationales, probing, and interpretability	Intrinsic interpretability design principles and architectures
9	Zhao, Zhou, Li, Tang, Dong, Hou, et al.	2026	Journal review article	Prompting, alignment, LLM evaluation, and preference optimization	Comprehensive review of pre-training, post-training, utilization, and evaluation of LLMs
10	Huang, Mo, Zhang, Li, Li, Zhang, et al.	2026	Journal review article	Prompting, alignment, LLM evaluation, and preference optimization	Multilingual LLM advances, language fairness, and accessibility
11	Ni, Liu, Wang, Lei, Zhao, Cheng, et al.	2025	Preprint	RAG, citation grounding, retrieval adequacy, and long-context methods	Trustworthy RAG framework spanning reliability, privacy, safety, fairness, explainability, and accountability
12	DeYoung, Jain, Rajani, Lehman, Xiong, Socher, et al.	2020	Conference paper	Explainable NLP, rationales, probing, and interpretability	Rationale evaluation and explanation benchmarks
13	Lei, Barzilay, Jaakkola	2016	Conference paper	Explainable NLP, rationales, probing, and interpretability	Rationale generation background
14	Camburu, Rocktaschel, Lukasiewicz, Blunsom	2018	Conference paper	Explainable NLP, rationales, probing, and interpretability	Explanation generation benchmark
15	Bastings, Aziz, Titov	2019	Conference paper	Explainable NLP, rationales, probing, and interpretability	Interpretable prediction
16	Jacovi, Goldberg	2020	Conference paper	Explainable NLP, rationales, probing, and interpretability	Definition of faithful explanation

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
17	Jain, Wallace	2019	Conference paper	Explainable NLP, rationales, probing, and interpretability	Attention–explanation limitation
18	Wiegreffe, Pinter	2019	Conference paper	Explainable NLP, rationales, probing, and interpretability	Nuanced attention–explanation debate
19	Chuang, Wang, Chang, Tang, Zhong, Yang, et al.	2026	Conference paper	Explainable NLP, rationales, probing, and interpretability	Model-agnostic evaluation and improvement of faithful natural-language explanations
20	Bilal, Ebert, Lin	2025	Preprint	Explainable NLP, rationales, probing, and interpretability	LLM-generated explanation and explainable AI evaluation challenges
21	Ribeiro, Singh, Guestrin	2016	Conference paper	Explainable NLP, rationales, probing, and interpretability	Local explanation background
22	Lundberg, Lee	2017	Conference paper	Explainable NLP, rationales, probing, and interpretability	Attribution evidence and its limitations
23	Doshi-Velez, Kim	2017	Preprint	Explainable NLP, rationales, probing, and interpretability	Evaluation rigor for explanation
24	Lipton	2018	Journal article	Explainable NLP, rationales, probing, and interpretability	Conceptual boundaries of interpretability
25	Rudin	2019	Journal article	Explainable NLP, rationales, probing, and interpretability	High-stakes explanation risk
26	Miller	2019	Journal article	Explainable NLP, rationales, probing, and interpretability	Human-centered explanation criteria
27	Hase, Bansal	2020	Conference paper	Explainable NLP, rationales, probing, and interpretability	Usefulness of explanations

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
28	Maynez, Narayan, Bohnet, McDonald	2020	Conference paper	Hallucination, factuality, and claim verification	Factuality and faithfulness in summarization
29	Kryscinski, McCann, Xiong, Socher	2020	Conference paper	Hallucination, factuality, and claim verification	Factual consistency evaluation
30	Huang, Yu, Ma, Zhong, Feng, Wang, et al.	2025	Journal article	Hallucination, factuality, and claim verification	Recent LLM hallucination taxonomy and open challenges
31	Manakul, Liusie, Gales	2023	Conference paper	Hallucination, factuality, and claim verification	Self-consistency hallucination detection
32	Lin, Hilton, Evans	2022	Conference paper	Hallucination, factuality, and claim verification	Truthfulness benchmark
33	Min, Krishna, Lyu, Lewis, Yih, Koh, et al.	2023	Conference paper	Hallucination, factuality, and claim verification	Atomic fact evaluation and its limits
34	Li, Cheng, Zhao, Nie, Wen	2023	Conference paper	Hallucination, factuality, and claim verification	Hallucination evaluation benchmark
35	Li, Khashabi, Khot, Sabharwal, Srikumar	2020	Conference paper	Causal language, uncertainty communication, and accountable human decision support	Underspecification and bias risk
36	Li, Zhang, Fan, Ding, Feng	2026	Conference paper	Hallucination, factuality, and claim verification	Distinguishes source fidelity from externally plausible factual correction
37	Scirè, Bejgu, Tedeschi, Ghonim, Martelli, Navigli	2026	Journal article	Hallucination, factuality, and claim verification	End-to-end factuality benchmark and evidence-based claim verification
38	Chen, Padmanabhan, Giyahchi, Wong, Akoglu	2026	Conference paper	Hallucination, factuality, and claim verification	Desiderata and evaluation design for LLM hallucination-detection benchmarks

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
39	Fang, Mackenzie	2026	Conference demonstration paper	Hallucination, factuality, and claim verification	Reproducible agentic fact search and claim-level verification
40	Thorne, Vlachos, Christodoulopoulos, Mittal	2018	Conference paper	Hallucination, factuality, and claim verification	Claim-verification background
41	Wadden, Lin, Lo, Wang, van Zuylen, Cohan, Hajishirzi	2020	Conference paper	Hallucination, factuality, and claim verification	Scientific claim verification
42	Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, et al.	2020	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	RAG as a grounding mechanism
43	Sarthi, Abdullah, Tuli, Khanna, Goldie, Manning	2024	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Hierarchical retrieval and its limits
44	Liu, Lin, Hewitt, Paranjape, Bevilacqua, Petroni, et al.	2024	Journal article	RAG, citation grounding, retrieval adequacy, and long-context methods	Long-context retrieval limits
45	Es, James, Espinosa-Anke, Schockaert	2024	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	RAG evaluation metrics
46	Saad-Falcon, Khattab, Potts, Zaharia	2024	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Automated RAG evaluation
47	Yoran, Wolfson, Ram, Berant	2024	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Robustness to irrelevant context
48	Chen, Lin, Han, Sun	2023	Preprint	RAG, citation grounding, retrieval adequacy, and long-context methods	RAG benchmarking

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
49	Zhang, Zhang	2025	Journal article	RAG, citation grounding, retrieval adequacy, and long-context methods	RAG hallucination mitigation and detection/correction framework
50	Karpukhin, Oguz, Min, Lewis, Wu, Edunov, et al.	2020	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Retrieval recall background
51	Izacard, Grave	2021	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Retrieval-conditioned generation
52	Borgeaud, Mensch, Hoffmann, Cai, Rutherford, Millican, et al.	2022	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Nonparametric memory and retrieval
53	Khattab, Santhanam, Li, Hall, Liang, Potts, et al.	2022	Preprint	RAG, citation grounding, retrieval adequacy, and long-context methods	Retrieval and prompting architecture
54	Gao, Yen, Yu, Chen	2023	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Citation grounding
55	Bohnet, Tran, Verga, Aharoni, Andor, Soares, et al.	2022	Preprint	RAG, citation grounding, retrieval adequacy, and long-context methods	Attribution and answer support
56	Rashkin, Nikolaev, Lamm, Aroyo, Collins, Das, et al.	2023	Journal article	RAG, citation grounding, retrieval adequacy, and long-context methods	Attribution and source support
57	Mallen, Asai, Zhong, Das, Khashabi, Hajishirzi	2023	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	When retrieval helps or fails

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
58	Trappolini, Cuconasu, Filice, Maarek, Silvestri	2026	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	LLM-oriented retrieval utility and distraction-aware evaluation
59	Murugaraj, Lamsiyah, Theobald	2026	Conference demonstration paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Diagnostic decomposition of RAG retrieval, completeness, faithfulness, and judge calibration
60	Choi, Guo, Fung, Wang	2026	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Retrieval-augmented citation attribution and analysis of LLM citation-judge failure
61	Ma, Chu, Fuhr	2026	Workshop paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Sentence-level citation faithfulness in retrieval-augmented medical QA
62	Chen, Li, Nie, Zhang, Ye, Zhao	2026	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Multi-turn report revision as an evaluation axis for deep-research agents
63	Pearl	2009	Book	Causal language, uncertainty communication, and accountable human decision support	Causal language boundaries
64	Miguel A. Hernan, James M. Robins	2020	Book	Causal language, uncertainty communication, and accountable human decision support	Causal claims and evidence standards
65	Feder, Keith, Manzoor, Pryzant, Sridhar, Wood-Doughty, et al.	2022	Journal article	Causal language, uncertainty communication, and accountable human decision support	Causal inference in NLP

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
66	Keith, Jensen, O'Connor	2020	Conference paper	Causal language, uncertainty communication, and accountable human decision support	Text-based causal inference background
67	Vig, Gehrmann, Belinkov, Qian, Nevo, Singer, Shieber	2020	Conference paper	Causal language, uncertainty communication, and accountable human decision support	Causal mediation analysis of gender bias in language models
68	van der Bles, van der Linden, Freeman, Mitchell, Galvao, Zaval, Spiegelhalter	2019	Journal article	Causal language, uncertainty communication, and accountable human decision support	Uncertainty communication principles
69	Spiegelhalter	2017	Journal article	Causal language, uncertainty communication, and accountable human decision support	Risk and uncertainty framing
70	Kunitomo-Jacquin, Marrese-Taylor, Fukuda, Hamasaki	2026	Conference paper	Causal language, uncertainty communication, and accountable human decision support	Evidential semantic entropy for LLM uncertainty quantification
71	Fadeeva, Rubashevskii, Piatrashyn, Vashurin, Dhuliawala, Shelmanov, et al.	2026	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Faithfulness-aware uncertainty quantification separating factuality from retrieved-evidence support
72	Ouyang, Wu, Jiang, Almeida, Wainwright, Mishkin, et al.	2022	Conference paper	Prompting, alignment, LLM evaluation, and preference optimization	Preference optimization and alignment

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
73	Bai, Kadavath, Kundu, Askell, Kernion, Jones, et al.	2022	Preprint	Prompting, alignment, LLM evaluation, and preference optimization	Constitutional alignment background
74	Bai, Jones, Ndousse, Askell, Chen, DasSarma, et al.	2022	Preprint	Prompting, alignment, LLM evaluation, and preference optimization	Helpful–harmless alignment
75	Nakano, Hilton, Balaji, Wu, Ouyang, Kim, et al.	2021	Preprint	Prompting, alignment, LLM evaluation, and preference optimization	Web-assisted answer generation
76	Liu, Shen, Zhang, Dolan, Carin, Chen	2022	Conference paper	Prompting, alignment, LLM evaluation, and preference optimization	Prompt sensitivity and context dependence
77	Thoppilan, De Freitas, Hall, Shazeer, Kulshreshtha, Cheng, et al.	2022	Preprint	Prompting, alignment, LLM evaluation, and preference optimization	LLM background
78	Touvron, Martin, Stone, Albert, Almahairi, Babaei, et al.	2023	Preprint	Prompting, alignment, LLM evaluation, and preference optimization	Foundation-model background
79	OpenAI, Achiam, Adler, Agarwal, Ahmad, Akkaya, et al.	2023	Technical report	Prompting, alignment, LLM evaluation, and preference optimization	LLM capabilities and limitations
80	Kaddour, Harris, Mozes, Bradley, Raileanu, McHardy	2023	Preprint	Prompting, alignment, LLM evaluation, and preference optimization	LLM challenges
81	Nori, King, McKinney, Carignan, Horvitz	2023	Preprint	Causal language, uncertainty communication, and accountable human decision support	High-stakes medical evaluation, calibration, and explanation example

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
82	Zheng, Chiang, Sheng, Zhuang, Wu, Zhuang, et al.	2023	Conference paper	Prompting, alignment, LLM evaluation, and preference optimization	LLM-as-a-judge limitations
83	Liu, Iter, Xu, Wang, Xu, Zhu	2023	Conference paper	Prompting, alignment, LLM evaluation, and preference optimization	LLM-based evaluation
84	Fu, Ng, Jiang, Liu	2023	Preprint	Prompting, alignment, LLM evaluation, and preference optimization	Prompted evaluation metrics
85	Wei, Wang, Schuurmans, Bosma, Ichter, Xia, et al.	2022	Conference paper	Prompting, alignment, LLM evaluation, and preference optimization	Generated reasoning traces
86	Wang, Wei, Schuurmans, Le, Chi, Narang, et al.	2023	Conference paper	Prompting, alignment, LLM evaluation, and preference optimization	Reasoning reliability background
87	Liu, Yuan, Fu, Jiang, Hayashi, Neubig	2023	Journal article	Prompting, alignment, LLM evaluation, and preference optimization	Prompting background
88	Mitchell, Wu, Zaldivar, Barnes, Vasserman, Hutchinson, et al.	2019	Conference paper	Causal language, uncertainty communication, and accountable human decision support	Model documentation and accountability
89	Gebru, Morgenstern, Vecchione, Vaughan, Wallach, Daume, Crawford	2021	Journal article	Causal language, uncertainty communication, and accountable human decision support	Dataset documentation
90	Bender, Gebru, McMillan-Major, Shmitchell	2021	Conference paper	Causal language, uncertainty communication, and accountable human decision support	LLM risk and accountability

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
91	Lai, Tan	2019	Conference paper	Causal language, uncertainty communication, and accountable human decision support	Human use of explanations
92	Kaur, Nori, Jenkins, Caruana, Wallach, Vaughan	2020	Conference paper	Causal language, uncertainty communication, and accountable human decision support	Human use of interpretability tools
93	Suresh, Gomez, Nam, Satyanarayan	2021	Conference paper	Causal language, uncertainty communication, and accountable human decision support	Stakeholder-specific accountability needs
94	Amershi, Weld, Vorvoreanu, Fourney, Nushi, Collisson, et al.	2019	Conference paper	Causal language, uncertainty communication, and accountable human decision support	Human–AI interaction principles
95	Buçinca, Malaya, Gajos	2021	Journal article	Causal language, uncertainty communication, and accountable human decision support	Overreliance and cognitive forcing
96	Parasuraman, Sheridan, Wickens	2000	Journal article	Causal language, uncertainty communication, and accountable human decision support	Human control and automation levels
97	Endsley	1995	Journal article	Causal language, uncertainty communication, and accountable human decision support	Situation awareness and decision-facing communication

(Continued)

Table A1 (continued)

No.	Source Short Label	Year	Venue Type	Primary Coding Category	Role in Taxonomy
98	Wang, He, Deng, Li, Li	2026	Journal article	Causal language, uncertainty communication, and accountable human decision support	Engineering generalizability case for NLP-assisted transfer learning and evidence-to-action constraints
99	Chae, Yeom, Park, Bae, Jang, Jin, et al.	2026	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Structure- and safety-aware statute-centric retrieval, citation, and abstention benchmark
100	Liu, Zhang, Ma, Deng, Zhu, Li et al.	2026	Conference survey paper	Causal language, uncertainty communication, and accountable human decision support	Legal-agent taxonomy, evaluation, verification, and accountability challenges
101	Papineni, Roukos, Ward, Zhu	2002	Conference paper	Prompting, alignment, LLM evaluation, and preference optimization	Limitations of surface-overlap metrics
102	Lin	2004	Conference paper	Prompting, alignment, LLM evaluation, and preference optimization	Limitations of summarization-overlap metrics
103	Zhang, Kishore, Wu, Weinberger, Artzi	2020	Conference paper	Prompting, alignment, LLM evaluation, and preference optimization	Semantic-similarity metric limitations
104	Reimers, Gurevych	2019	Conference paper	RAG, citation grounding, retrieval adequacy, and long-context methods	Semantic similarity and retrieval background

Note: AI: artificial intelligence, LLM: large language model, NLP: natural language processing, RAG: retrieval-augmented generation.

References

1. Danilevsky M, Qian K, Aharonov R, Katsis Y, Kawas B, Sen P. A survey of the state of explainable AI for natural language processing. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing; 2020 Dec 4–7; Suzhou, China. p. 447–59. doi:10.18653/v1/2020.aacl-main.46.
2. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. arXiv:2312.10997. 2023. doi:10.48550/arXiv.2312.10997.

3. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1–38. doi:10.1145/3571730.
4. Kotonya N, Toni F. Explainable automated fact-checking: a survey. In: Proceedings of the 28th International Conference on Computational Linguistics; 2020 Dec 8–13; Barcelona, Spain. p. 5430–43. doi:10.18653/v1/2020.coling-main.474.
5. Lyu Q, Apidianaki M, Callison-Burch C. Towards faithful model explanation in NLP: a survey. Comput Linguist. 2024;50(2):657–723. doi:10.1162/coli_a_00511.
6. Qin L, Chen Q, Feng X, Wu Y, Zhang Y, Li Y, et al. Large language models meet NLP: a survey. Front Comput Sci. 2026;20(11):2011361. doi:10.1007/s11704-025-50472-3.
7. Javed H, Shah B, Ali F, Kwak D. Large language models in NLP: evolution, architectural trends, and open challenges. J Big Data. 2026;13(1):95. doi:10.1186/s40537-026-01429-1.
8. Gao Y, Meng Q, Zhou Y, Pan L. Towards intrinsic interpretability of large language models: a survey of design principles and architectures. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2026 Jul 2–7; San Diego, CA, USA. p. 34741–54. doi:10.18653/v1/2026.acl-long.1605.
9. Zhao WX, Zhou K, Li J, Tang T, Dong Z, Hou Y, et al. A survey of large language models. Front Comput Sci. 2026;20(12):2012627. doi:10.1007/s11704-026-60308-3.
10. Huang K, Mo F, Zhang X, Li H, Li Y, Zhang Y, et al. A survey on large language models with multilingualism: recent advances and new frontiers. Artif Intell Rev. 2026;59(6):146. doi:10.1007/s10462-026-11534-5.
11. Ni B, Liu Z, Wang L, Lei Y, Zhao Y, Cheng X, et al. Towards trustworthy retrieval augmented generation for large language models: a survey. arXiv:2502.06872. 2025. doi:10.48550/arXiv.2502.06872.
12. DeYoung J, Jain S, Rajani NF, Lehman E, Xiong C, Socher R, et al. ERASER: a benchmark to evaluate rationalized NLP models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 4443–58. doi:10.18653/v1/2020.acl-main.408.
13. Lei T, Barzilay R, Jaakkola T. Rationalizing neural predictions. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016 Nov 1–5; Austin, TX, USA. p. 107–17. doi:10.18653/v1/D16-1011.
14. Camburu OM, Rockt?schel T, Lukasiewicz T, Blunsom P. e-SNLI: natural language inference with natural language explanations. In: Proceedings of the 32nd Annual Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montréal, QC, Canada. p. 9539–49.
15. Bastings J, Aziz W, Titov I. Interpretable neural predictions with differentiable binary variables. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. p. 2963–77. doi:10.18653/v1/P19-1284.
16. Jacovi A, Goldberg Y. Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 4198–205. doi:10.18653/v1/2020.acl-main.386.
17. Jain S, Wallace BC. Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minneapolis, MN, USA. p. 3543–56. doi:10.18653/v1/N19-1357.
18. Wiegreffe S, Pinter Y. Attention is not not explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China. p. 11–20. doi:10.18653/v1/D19-1002.
19. Chuang YN, Wang G, Chang CY, Tang R, Zhong S, Yang F, et al. FaithLM: towards faithful explanations for large language models. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); 2026 Mar 24–29; Rabat, Morocco. p. 3802–24. doi:10.18653/v1/2026.eacl-long.177.
20. Bilal A, Ebert D, Lin B. LLMs for explainable AI: a comprehensive survey. arXiv:2504.00125. 2025. doi:10.48550/arXiv.2504.00125.

21. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. p. 1135–44. doi:10.1145/2939672.2939778.
22. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 4765–74.
23. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608. 2017. doi:10.48550/arXiv.1702.08608.
24. Lipton ZC. The mythos of model interpretability. Commun ACM. 2018;61(10):36–43. doi:10.1145/3233231.
25. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. doi:10.1038/s42256-019-0048-x.
26. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell. 2019;267(2):1–38. doi:10.1016/j.artint.2018.07.007.
27. Hase P, Bansal M. Evaluating explainable AI: which algorithmic explanations help users predict model behavior? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 5540–52. doi:10.18653/v1/2020.acl-main.491.
28. Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 1906–19. doi:10.18653/v1/2020.acl-main.173.
29. Kryscinski W, McCann B, Xiong C, Socher R. Evaluating the factual consistency of abstractive text summarization. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; 2020 Nov 16–20; Online. p. 9332–46. doi:10.18653/v1/2020.emnlp-main.750.
30. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. 2025;43(2):1–55. doi:10.1145/3703155.
31. Manakul P, Liusie A, Gales MJF. SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. p. 9004–17. doi:10.18653/v1/2023.emnlp-main.557.
32. Lin S, Hilton J, Evans O. TruthfulQA: measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022 May 22–27; Dublin, Ireland. p. 3214–52. doi:10.18653/v1/2022.acl-long.229.
33. Min S, Krishna K, Lyu X, Lewis M, Yih W, Koh PW, et al. FActScore: fine-grained atomic evaluation of factual precision in long form text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. p. 12076–100. doi:10.18653/v1/2023.emnlp-main.741.
34. Li J, Cheng X, Zhao WC, Nie JY, Wen JR. HaluEval: a large-scale hallucination evaluation benchmark for large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. p. 6449–64. doi:10.18653/v1/2023.emnlp-main.397.
35. Li T, Khashabi D, Khot T, Sabharwal A, Srikumar V. UNQOVERing stereotyping biases via underspecified questions. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings; 2020 Nov 16–20; Online. p. 3475–89. doi:10.18653/v1/2020.findings-emnlp.311.
36. Li M, Zhang H, Fan H, Ding J, Feng Y. Harmful factuality: LLMs correcting what they shouldn't. In: Proceedings of the Findings of the Association for Computational Linguistics: EACL 2026; 2026 Mar 24–29; Rabat, Morocco. p. 896–912. doi:10.18653/v1/2026.findings-eacl.46.
37. Scirè A, Bejgu AS, Tedeschi S, Ghonim K, Martelli F, Navigli R. Truth or mirage? Towards end-to-end factuality evaluation with LLM-Oasis. Comput Linguist. 2026;52(1):1–41. doi:10.1162/coli.a.575.
38. Chen W, Padmanabhan V, Giyahchi T, Wong E, Akoglu L. Rethinking evaluation for LLM hallucination detection: a desiderata, a new RAG-based benchmark, new insights. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2026 Jul 2–7; San Diego, CA, USA. p. 14912–31. doi:10.18653/v1/2026.acl-long.680.

39. Fang M, Mackenzie H. FactSearch: an interactive agentic fact search system for verifying large language model outputs. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations); 2026 Jul 2–7; San Diego, CA, USA. p. 367–73. doi:10.18653/v1/2026.acl-demo.36.
40. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a large-scale dataset for fact extraction and verification. In: Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018 Jun 1–6; New Orleans, LA, USA. p. 809–19. doi:10.18653/v1/N18-1074.
41. Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or fiction: verifying scientific claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; 2020 Nov 16–20; Online. p. 7534–50. doi:10.18653/v1/2020.emnlp-main.609.
42. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th Annual Conference on Neural Information Processing Systems; 2020 Dec 6–12; Online. p. 9459–74.
43. Sarthi P, Abdullah S, Tuli A, Khanna A, Goldie A, Manning CD. RAPTOR: recursive abstractive processing for tree-organized retrieval. In: Proceedings of the 12th International Conference on Learning Representations; 2024 May 7–11; Vienna, Austria.
44. Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, et al. Lost in the middle: how language models use long contexts. Trans Assoc Comput Linguist. 2024;12(5):157–73. doi:10.1162/tacl_a_00638.
45. Es S, James J, Espinosa-Anke L, Schockaert S. RAGAS: automated evaluation of retrieval augmented generation. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics; 2024 Mar 17–22; St Julian's, Malta. p. 150–8. doi:10.18653/v1/2024.eacl-demo.16.
46. Saad-Falcon J, Khattab O, Potts C, Zaharia M. ARES: an automated evaluation framework for retrieval-augmented generation systems. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2024 Jun 16–21; Mexico City, Mexico. p. 338–54. doi:10.18653/v1/2024.naacl-long.20.
47. Yoran O, Wolfson T, Ram O, Berant J. Making retrieval-augmented language models robust to irrelevant context. In: Proceedings of the 12th International Conference on Learning Representations; 2024 May 7–11; Vienna, Austria.
48. Chen J, Lin H, Han X, Sun L. Benchmarking large language models in retrieval-augmented generation. arXiv:2309.01431. 2023. doi:10.48550/arXiv.2309.01431.
49. Zhang W, Zhang J. Hallucination mitigation for retrieval-augmented large language models: a review. Mathematics. 2025;13(5):856. doi:10.3390/math13050856.
50. Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; 2020 Nov 16–20; Online. p. 6769–81. doi:10.18653/v1/2020.emnlp-main.550.
51. Izacard G, Grave E. Leveraging passage retrieval with generative models for open-domain question answering. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics; 2021 Apr 19–23; Online. p. 874–80. doi:10.18653/v1/2021.eacl-main.74.
52. Borgeaud S, Mensch A, Hoffmann J, Cai T, Rutherford E, Millican K, et al. Improving language models by retrieving from trillions of tokens. In: Proceedings of the 39th International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. p. 2206–40.
53. Khattab O, Santhanam K, Li X, Hall D, Liang P, Potts C, et al. Demonstrate-search-predict: composing retrieval and language models for knowledge-intensive NLP. arXiv:2212.14024. 2022. doi:10.48550/arXiv.2212.14024.
54. Gao T, Yen H, Yu J, Chen D. Enabling large language models to generate text with citations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. p. 6465–88. doi:10.18653/v1/2023.emnlp-main.398.
55. Bohnet B, Tran VQ, Verga P, Aharoni R, Andor D, Soares L, et al. Attributed question answering: evaluation and modeling for attributed large language models. arXiv:2212.08037. 2022. doi:10.48550/arXiv.2212.08037.
56. Rashkin H, Nikolaev V, Lamm M, Aroyo L, Collins M, Das D, et al. Measuring attribution in natural language generation models. Comput Linguist. 2023;49(4):777–840. doi:10.1162/coli_a_00486.

57. Mallen A, Asai A, Zhong V, Das R, Khashabi D, Hajishirzi H. When not to trust language models: investigating effectiveness of parametric and non-parametric memories. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023 Jul 9–14; Toronto, ON, Canada. p. 9802–22. doi:10.18653/v1/2023.acl-long.546.
58. Trappolini G, Cuconasu F, Filice S, Maarek Y, Silvestri F. Redefining retrieval evaluation in the era of LLMs. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); 2026 Mar 24–29; Rabat, Morocco. p. 8359–75. doi:10.18653/v1/2026.eacl-long.391.
59. Murugaraj K, Lamsiyah S, Theobald M. RAGVUE: a diagnostic view for explainable and automated evaluation of retrieval-augmented generation. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations); 2026 Mar 24–29; Rabat, Morocco. p. 512–26. doi:10.18653/v1/2026.eacl-demo.35.
60. Choi YM, Guo X, Fung YR, Wang Q. CiteGuard: faithful citation attribution for LLMs via retrieval-augmented validation. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2026 Jul 2–7; San Diego, CA, USA. p. 6241–57. doi:10.18653/v1/2026.acl-long.282.
61. Ma Y, Chu B, Fuhr N. VERICITE: evaluating sentence-level citation faithfulness in retrieval-augmented medical question answering. In: Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026); 2026 Jul 3–4; San Diego, CA, USA. p. 753–9. doi:10.18653/v1/2026.bionlp-1.62.
62. Chen B, Li B, Nie P, Zhang Y, Ye X, Zhao C. Beyond single-shot writing: deep research agents are unreliable at multi-turn report revision. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2026 Jul 2–7; San Diego, CA, USA. p. 13325–56. doi:10.18653/v1/2026.acl-long.609.
63. Pearl J. Causality: models, reasoning, and inference. 2nd ed. Cambridge, UK: Cambridge University Press; 2009. doi:10.1017/cbo9780511803161.
64. Hernán MA, Robins JM. Causal inference: what if. Boca Raton, FL, USA: Chapman & Hall/CRC; 2020.
65. Feder A, Keith KA, Manzoor E, Pryzant R, Sridhar D, Wood-Doughty Z, et al. Causal inference in natural language processing: estimation, prediction, interpretation and beyond. Trans Assoc Comput Linguist. 2022;10(1):1138–58. doi:10.1162/tacl_a_00511.
66. Keith KA, Jensen D, O'Connor B. Text and causal inference: a review of using text to remove confounding from causal estimates. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 5332–44. doi:10.18653/v1/2020.acl-main.474.
67. Vig J, Gehrmann S, Belinkov Y, Qian S, Nevo D, Singer Y, et al. Investigating gender bias in language models using causal mediation analysis. In: Proceedings of the 34th Annual Conference on Neural Information Processing Systems; 2020 Dec 6–12; Online. p. 12388–401.
68. van der Bles AM, van der Linden S, Freeman ALJ, Mitchell J, Galvao AB, Zaval L, et al. Communicating uncertainty about facts, numbers and science. R Soc Open Sci. 2019;6(5):181870. doi:10.1098/rsos.181870.
69. Spiegelhalter D. Risk and uncertainty communication. Annu Rev Stat Appl. 2017;4(1):31–60. doi:10.1146/annurev-statistics-010814-020148.
70. Kunitomo-Jacquin L, Marrese-Taylor E, Fukuda K, Hamasaki M. Evidential semantic entropy for LLM uncertainty quantification. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); 2026 Mar 24–29; Rabat, Morocco. p. 7107–22. doi:10.18653/v1/2026.eacl-long.334.
71. Fadeeva E, Rubashevskii A, Piatrashyn D, Vashurin R, Dhuliawala S, Shelmanov A, et al. Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval-augmented generation. In: Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; 2026 Jul 2–7; San Diego, CA, USA. p. 6814–36. doi:10.18653/v1/2026.findings-acl.338.
72. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th Annual Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 27730–44.

73. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: harmlessness from AI feedback. arXiv:2212.08073. 2022. doi:10.48550/arXiv.2212.08073.
74. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862. 2022. doi:10.48550/arXiv.2204.05862.
75. Nakano R, Hilton J, Balaji S, Wu J, Ouyang L, Kim C, et al. WebGPT: browser-assisted question-answering with human feedback. arXiv:2112.09332. 2021. doi:10.48550/arXiv.2112.09332.
76. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What makes good in-context examples for GPT-3? In: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; 2022 May 26–27; Dublin, Ireland. p. 100–14. doi:10.18653/v1/2022.deelio-1.10.
77. Thoppilan R, de Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng HT, et al. LaMDA: language models for dialog applications. arXiv:2201.08239. 2022. doi:10.48550/arXiv.2201.08239.
78. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. arXiv:2307.09288. 2023. doi:10.48550/arXiv.2307.09288.
79. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. arXiv:2303.08774. 2023. doi:10.48550/arXiv.2303.08774.
80. Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. arXiv:2307.10169. 2023. doi:10.48550/arXiv.2307.10169.
81. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375. 2023. doi:10.48550/arXiv.2303.13375.
82. Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: Proceedings of the 37th Annual Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. p. 46595–623. doi:10.52202/075280-2020.
83. Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. p. 2511–22. doi:10.18653/v1/2023.emnlp-main.153.
84. Fu J, Ng SK, Jiang Z, Liu P. GPTScore: evaluate as you desire. arXiv:2302.04166. 2023. doi:10.48550/arXiv.2302.04166.
85. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th Annual Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 24824–37. doi:10.52202/068431-1800.
86. Wang X, Wei J, Schuurmans D, Le QV, Chi EH, Narang S, et al. Self-consistency improves chain-of-thought reasoning in language models. In: Proceedings of the 11th International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda. p. 1–24.
87. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv. 2023;55(9):1–35. doi:10.1145/3560815.
88. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency; 2019 Jan 29–31; Atlanta, GA, USA. p. 220–9. doi:10.1145/3287560.3287596.
89. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86–92. doi:10.1145/3458723.
90. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 Mar 3–10; Online. p. 610–23. doi:10.1145/3442188.3445922.
91. Lai V, Tan C. On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In: Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency; 2019 Jan 29–31; Atlanta, GA, USA. p. 29–38. doi:10.1145/3287560.3287590.

92. Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Vaughan JW. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020 Apr 25–30; Honolulu, HI, USA. p. 1–14. doi:10.1145/3313831.3376219.
93. Suresh H, Gomez SR, Nam KK, Satyanarayan A. Beyond expertise and roles: a framework to characterize the stakeholders of interpretable machine learning and their needs. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021 May 8–13; Yokohama, Japan. p. 1–16. doi:10.1145/3411764.3445088.
94. Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, et al. Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; 2019 May 4–9; Glasgow, UK. p. 1–13. doi:10.1145/3290605.3300233.
95. Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc ACM Hum-Comput Interact. 2021;5(CSCW1):1–21. doi:10.1145/3449287.
96. Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern Part A Syst Hum. 2000;30(3):286–97. doi:10.1109/3468.844354.
97. Endsley MR. Toward a theory of situation awareness in dynamic systems. Hum Factors. 1995;37(1):32–64. doi:10.1518/001872095779049543.
98. Wang S, He J, Deng W, Li Y, Li B. Enhanced creep strain prediction in diverse adhesively bonded joints using natural language processing-assisted transfer learning. Eng Appl Artif Intell. 2026;167:113872. doi:10.1016/j.engappai.2026.113872.
99. Chae K, Yeom J, Park J, Bae S, Jang I, Jin H, et al. Evaluating structure-aware retrieval and safety in statute-centric legal QA. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2026 Jul 2–7; San Diego, CA, USA. p. 45553–73. doi:10.18653/v1/2026.acl-long.2112.
100. Liu S, Zhang R, Ma R, Deng Y, Zhu L, Li J, et al. LLM agents in law: taxonomy, applications, and challenges. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2026 Jul 2–7; San Diego, CA, USA. p. 15768–92. doi:10.18653/v1/2026.acl-long.718.
101. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002 Jul 7–12; Philadelphia, PA, USA. p. 311–8. doi:10.3115/1073083.1073135.
102. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out; 2004 Jul 25–26; Barcelona, Spain. p. 74–81.
103. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. In: Proceedings of the 8th International Conference on Learning Representations; 2020 Apr 26–May 1; Online.
104. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-network. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China. p. 3982–92. doi:10.18653/v1/D19-1410.