Passage Influence in Genetic Evidence Summaries

Elijah King, Caleb Wright, Nathaniel Scott

Abstract

Passage Influence in Genetic Evidence Summaries addresses a narrow but consequential interpretive problem: one highly ranked passage controlling a broad claim. This literature-based conceptual analysis does not report a new experiment. The focal evidence concerns how exploratory cross-trait analysis might nominate candidate regions shared by hearing loss and age-related disease and connect them to inner-ear contexts. Attribution asks what evidence influenced a statement and whether that evidence supports it. Those are different tests. Remove and replace passages while monitoring individual claims. Pair influence analysis with qualified support judgments, retaining failures in either dimension. The proposed comparison concerns candidate regions, source snippets, generated interpretation; it would test whether the specific failure scenario is detected without erasing legitimate uncertainty. An integrated account should preserve the distinction between candidate regions and cellular annotations and any stronger downstream decision.

The failure scenario

Candidate regions and generated interpretation occupy different positions in the evidence chain. Their relationship is the starting point for examining attribution fidelity. Consider the possibility of one highly ranked passage controlling a broad claim. This is a proposed stress scenario, not an observed event reported by the cited studies. Its purpose is to expose which evidence contract would be violated even if the final output looked plausible. For candidate regions and cellular annotations, the relevant claim must identify its population or artifact class, its time window, and the action it is supposed to support. A failure at any of these levels may persist even after a local technical improvement. The analysis therefore treats published methods as bounded evidence and asks how their limitations interact with this particular decision. An operational analogy is useful only if it keeps those limitations visible.

Attribution asks what evidence influenced a statement and whether that evidence supports it. Those are different tests. A model may faithfully rely on an irrelevant passage, while a correct statement may receive a decorative citation unrelated to its generation. In this article, the relevant contrast is candidate regions, source snippets, generated interpretation. The proposed counterexample of one highly ranked passage controlling a broad claim puts that contrast under pressure. Remove and replace passages while monitoring individual claims. Pair influence analysis with qualified support judgments, retaining failures in either dimension. A favorable finding would show that the intended interpretation of generated interpretation remains defensible when candidate regions is perturbed under the specified conditions. An unfavorable finding should narrow the claim rather than be discarded as an inconvenient case.

The transition at issue is specific to exploratory hearing and age-related genetic synthesis. It concerns candidate regions, source snippets, generated interpretation, and the scenario of one highly ranked passage controlling a broad claim shows why a locally correct result can still become a misleading integrated claim. The input record should retain enough information to decide which part of that scenario violates the intended contract. That includes the original source type, measurement or artifact status, and the reason the proposed interpretation exceeds its support. The evaluator would compare the system's response with a qualified judgment made without seeing that response. Disagreement should be classified rather than erased by rewriting the prompt. A model may faithfully rely on an irrelevant passage, while a correct statement may receive a decorative citation unrelated to its generation. A useful negative result would show that the interface cannot preserve this distinction under realistic use; a useful positive result would demonstrate

preservation on independent cases rather than merely on the example used to develop the rule. The primary source remains relevant through its particular contribution: Its stated scope separates candidate nomination and cellular plausibility from confirmed causal mechanisms. That contribution should be retained even if the integrated workflow fails. Conversely, successful integration would not remove the source's limitations or turn a proposed extension into an already observed result. This asymmetric interpretation protects both the literature and the user from an unjustified all-or-nothing verdict.

Methods that constrain the claim

Retrieval-augmented generation makes external evidence part of the generation process (Lewis et al. 2020). Making documents available to a generator changes its information access but does not guarantee that a statement is supported. A useful analysis separates document selection from generation conditioned on those documents. Missing evidence, irrelevant context, and selective quotation create different failures. The proposed workflow should preserve retrieved passages and evaluate claim support independently of fluency or the mere presence of citations. Within exploratory hearing and age-related genetic synthesis, this distinction is relevant to attribution fidelity. The proposed failure scenario is one highly ranked passage controlling a broad claim; its evaluation must preserve the original source's methodological scope.

Dense retrieval provides semantic matching but must be compared with strong lexical baselines (Karpukhin et al. 2020). Semantic retrieval can connect questions and passages whose wording differs. That advantage should be tested against cases where wording differs because the underlying population or mechanism differs. Retrieval accuracy is not identical to answer accuracy, and neither automatically establishes evidence eligibility. A good evaluation retains the ranked candidates so later errors can be localized to retrieval rather than attributed to the generator by default. This is methodological support for examining attribution fidelity, not a claim that the source studied exploratory hearing and age-related genetic synthesis. A separate application-specific test is required at the interface described in this review.

Replication and transparent workflows are prerequisites for reliable biomedical claims (National Academies 2019). Reproducing a calculation and independently replicating a finding are different achievements. Recoverable data and code allow one to inspect the first; independent observations and design choices inform the second. A review should avoid calling several reanalyses of one dataset separate replications. Clear reporting of analysis changes also makes a failed replication informative instead of allowing it to disappear as an implementation dispute. An implementation should expose this limitation to reviewers of candidate regions and cellular annotations. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

FAIR principles support reusable cohorts, signals, and analysis metadata (Wilkinson et al. 2016). Reusable evidence needs more than a downloadable file. Identifiers, metadata, provenance, and meaningful access conditions help another researcher determine what an artifact contains and whether it can answer a new question. Reuse does not mean universal permission or universal scientific comparability. The information needed to judge those boundaries should travel with the artifact rather than remain in private workflow knowledge. For the comparison involving candidate regions, source snippets, generated interpretation, the source supplies an analytical obligation. The review's proposed application remains a hypothesis, with the deliberately difficult scenario serving as a check on scope inflation.

Causal inference requires a clearly specified intervention, time order, and exchangeability assumptions (Hernán and Robins 2020). A prediction can be accurate without identifying what an intervention would change. A useful causal analysis begins by defining the hypothetical comparison, eligibility, and time zero. Its assumptions concern the process generating exposure and observation, not merely the flexibility of a

fitted model. In this review, causal vocabulary is appropriate only when that comparison can be specified and defended. This is methodological support for examining attribution fidelity, not a claim that the source studied exploratory hearing and age-related genetic synthesis. A separate application-specific test is required at the interface described in this review.

Epidemiologic design makes confounding, selection, and measurement error part of interpretation (Rothman, Greenland, and Lash 2008). Confounding, selection, and measurement error are distinct problems and need not move an estimate in the same direction. An analysis should describe them before deciding which adjustment is adequate. A larger dataset can reduce random error while leaving systematic error almost untouched. Precision should therefore be interpreted alongside the quality of the design and the eligibility of its population. This is methodological support for examining attribution fidelity, not a claim that the source studied exploratory hearing and age-related genetic synthesis. A separate application-specific test is required at the interface described in this review.

False-discovery control is central when many molecular features are screened (Benjamini and Hochberg 1995). Multiplicity arises from the family of questions actually searched. Controlling false discoveries within a named family does not automatically cover later tissue filters, alternative endpoints, or narrative choices. The proposed audit should retain exploratory branches so the family is visible. Independent confirmation then addresses a different question from whether a discovery threshold was correctly applied during screening. An implementation should expose this limitation to reviewers of candidate regions and cellular annotations. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

Medical machine learning must fit data provenance, care pathways, and prospective evaluation (Rajkomar, Dean, and Kohane 2019). The clinical relevance of a model depends on how its output changes a care process. Development data can contain information that becomes available only after the intended decision. A review must identify that availability boundary before comparing algorithms. Prospective use also changes who is measured and who is treated, so evaluation of a frozen historical dataset cannot establish the entire future workflow. This is methodological support for examining attribution fidelity, not a claim that the source studied exploratory hearing and age-related genetic synthesis. A separate application-specific test is required at the interface described in this review.

Responsible health AI connects prediction with workflow, monitoring, and patient impact (Wiens et al. 2019). The point at which a model reaches practice is not the end of evaluation. New populations, changing measurements, and feedback can alter failure modes. A responsible deployment needs a route for recognizing deterioration and a person authorized to respond. Monitoring should be linked to feasible corrective actions; an alert that cannot change the workflow provides accountability only in a very limited sense. Within exploratory hearing and age-related genetic synthesis, this distinction is relevant to attribution fidelity. The proposed failure scenario is one highly ranked passage controlling a broad claim; its evaluation must preserve the original source's methodological scope.

Policy distillation connects compression with sequential decision behavior (Rusu et al. 2015). Compressing a policy concerns sequential behavior rather than one isolated prediction. A small local mismatch can compound across actions, while a locally similar policy can explore differently. Evaluation should retain trajectories and consequential state transitions. The review treats policy transfer as a potential operational tool whose task-level contract still needs checking in the destination environment. Within exploratory hearing and age-related genetic synthesis, this distinction is relevant to attribution fidelity. The proposed failure scenario is one highly ranked passage controlling a broad claim; its evaluation must preserve the original source's methodological scope.

Reading the literature by evidence type

A useful boundary case is supplied by Exploratory Cross-Trait Genetic Analysis of Hearing Loss and Age-Related Diseases Nominates Cardiometabolic Candidate Regions and Inner-Ear Cellular Contexts, concerned with how exploratory cross-trait analysis might nominate candidate regions shared by hearing loss and age-related disease and connect them to inner-ear contexts (Xiao Yu et al. 2026b). Its methodological basis is the supplied title identifies exploratory cross-trait genetic analysis and cellular-context nomination; an authoritative abstract was not retrieved. The accessible evidence is specific: only the supplied bibliographic scope supports discussion here; specific cohorts, variants, effect sizes, datasets, and analysis algorithms remain unverified. Its stated scope separates candidate nomination and cellular plausibility from confirmed causal mechanisms. Without accessible methods and results, the work cannot substantiate named loci, causal direction, treatment targets, or clinical risk improvements in this review. The source supplies one part of the comparison involving candidate regions, source snippets, generated interpretation. Its evidence should retain its original unit and scope; the present synthesis proposes a connection without claiming that the paper directly tested this integrated workflow.

A useful boundary case is supplied by BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models, concerned with how reinforcement learning for program repair can overcome sparse execution feedback and coarse sequence-level credit (Yuanhao Li et al. 2026a). Its methodological basis is a three-stage pipeline combining execution-verified supervised traces, sequence- and line-level reward models, and PPO with reward redistribution to critical edit regions. The accessible evidence is specific: evaluation spans SWE-bench Verified, Defects4J transfer, HumanEval-Java, and QuixBugs, with reported gains in both repository repair and cross-language generalization. Line-level credit assignment makes test execution informative at the granularity where patches actually change behavior. Benchmark success does not by itself establish maintainability, specification fidelity, or robustness to incomplete and flaky tests. This source enters the present review as a methodological comparison for evidence processing, not as evidence about disease biology, patient outcomes, or clinical efficacy. The transferable question is whether candidate regions remains correctly interpreted after a computational choice involving generated interpretation. Its task-specific validation cannot be borrowed to validate the clinical claim.

The relevant question in An Oxidative Stress Biomarker for Noise-Associated Stroke: Evidence from Human and Mice is how oxidative-stress-related evidence can link noise exposure and stroke across human and experimental settings (CongCong Wu et al. 2026). Its methodological basis is gene-expression screening with machine-learning approaches, longitudinal human analysis, a four-week noise-exposed mouse experiment, and NHANES observational analysis. The accessible evidence is specific: the abstract identifies CCL3 as a candidate and reports experimental gene checks and associations of oxidative-balance scores in noise-exposed participants. The paper combines several evidence scales rather than relying on one observational correlation. Triangulation does not eliminate confounding or species differences, and oxidative-balance associations are not evidence that a particular dietary intervention prevents stroke. The source supplies one part of the comparison involving candidate regions, source snippets, generated interpretation. Its evidence should retain its original unit and scope; the present synthesis proposes a connection without claiming that the paper directly tested this integrated workflow.

Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs asks how fine-tuned language models remain reliable when retrieved context is noisy or irrelevant (Yinghao. “Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs.” *2025 6th International Conference on Artificial Intelligence Sang 2025). Its methodological basis is the supplied IEEE conference title frames retrieval-noise robustness in fine-tuned

LLMs; detailed methods were not accessible. The accessible evidence is specific: the official IEEE document record confirms the bibliographic item but does not establish a verified noise protocol or quantitative result in the accessible material. Its stated problem makes the retrieval interface, rather than only model weights, a locus of robustness evaluation. No specific training method, benchmark, noise rate, or effect size is attributed without the full technical record. This source enters the present review as a methodological comparison for evidence processing, not as evidence about disease biology, patient outcomes, or clinical efficacy. The transferable question is whether candidate regions remains correctly interpreted after a computational choice involving generated interpretation. Its task-specific validation cannot be borrowed to validate the clinical claim.

The sources should be compared through the roles they play, not merged into a single performance claim. Exploratory Cross-Trait Genetic Analysis of Hearing Loss and Age-Related Diseases Nominates Cardiometabolic Candidate Regions and Inner-Ear Cellular Contexts addresses how exploratory cross-trait analysis might nominate candidate regions shared by hearing loss and age-related disease and connect them to inner-ear contexts (Xiao Yu et al. 2026b). By contrast, Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs addresses how fine-tuned language models remain reliable when retrieved context is noisy or irrelevant (Yinghao. “Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs.” *2025 6th International Conference on Artificial Intelligence Sang 2025). That difference matters for attribution fidelity: the evidence concerning candidate regions is not automatically evidence concerning generated interpretation. A successful software operation, a plausible mechanistic account, and a validated outcome are different achievements. Their integration may improve traceability, but it cannot create an unmeasured link. A coherent review should identify where the sources complement one another and where their scopes remain incompatible. Here the proposed failure scenario is one highly ranked passage controlling a broad claim. The analytical value of that scenario is that it can occur without requiring any cited paper to be wrong: an otherwise valid local finding can be misused at a later interface. Preventing that misuse requires evidence about the interface itself, not simply a larger collection of local results.

Research design and stopping conditions

A research agenda should address the specific negative case of one highly ranked passage controlling a broad claim. Remove and replace passages while monitoring individual claims. Pair influence analysis with qualified support judgments, retaining failures in either dimension. These procedures are proposed, not completed. The case set would include the scenario of one highly ranked passage controlling a broad claim, with its validity judged independently of the system's response. The relevant comparison axes are candidate regions, source snippets, generated interpretation.

An evidence-linking protocol would use a set of individually judged claims, eligible supporting passages, plausible but ineligible passages, and genuine counterevidence. The evaluation would retain rankings and intermediate interpretations. Source deletion and substitution would test whether support changes coherently. Fluency, citation presence, source influence, and actual support would be recorded separately, so an answer cannot receive full credit simply for sounding well documented.

For the exploratory genetic source, obtaining and inspecting the primary methods is a prerequisite to stronger attribution. The proposed review can presently test evidence-role preservation, but it cannot validate unverified genetic regions, sample construction, or tissue mappings. Any later biological protocol would need to be designed around the actual reported candidates rather than invented examples assigned to the publication.

What remains unverified

Access limits are material for Exploratory Cross-Trait Genetic Analysis of Hearing Loss and Age-Related Diseases Nominates Cardiometabolic Candidate Regions and Inner-Ear Cellular Contexts, Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs. Where only bibliographic scope or a limited abstract is available, the discussion does not assign unverified datasets, algorithms, effect sizes, or clinical recommendations. This review is a bounded conceptual synthesis rather than a systematic review with an exhaustive search protocol. It cannot establish the effectiveness of the proposed integrated workflow. Its specific contribution is to articulate how attribution fidelity changes the interpretation of candidate regions, source snippets, generated interpretation. Technical sources support computational comparisons only; they do not validate biological mechanisms or patient benefit. Reviews and correspondence retain their publication roles and are not counted as new experiments. Proposed stress cases are illustrative, not observed failures attributed to the papers. Future work should therefore test the weakest inference described here, preserve contrary results, and disclose the cost of checking rather than treating additional automation as intrinsically beneficial.

Conclusion

The most valuable next result would be a discriminating test of attribution fidelity, including a realistic way for the preferred approach to fail. Remove and replace passages while monitoring individual claims. Pair influence analysis with qualified support judgments, retaining failures in either dimension. The primary source remains bounded by its own design: Without accessible methods and results, the work cannot substantiate named loci, causal direction, treatment targets, or clinical risk improvements in this review. A trustworthy workflow would preserve the evidence source, the assumptions of the transition, and the right to stop when support is insufficient. The proposed evaluation agenda gives those obligations observable tests while withholding claims of completed validation.

References

1. Yu, Xiao, et al. "Exploratory Cross-Trait Genetic Analysis of Hearing Loss and Age-Related Diseases Nominates Cardiometabolic Candidate Regions and Inner-Ear Cellular Contexts." *Intelligent Translational Medicine*, 2026, https://doi.org/10.1016/j.itm.2026.100002.

2. Li, Y., Wang, H., Shang, X., Tang, X., Cao, Y., & Chen, X. (2026). BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models. arXiv preprint arXiv:2605.09134.

3. Wu, CongCong, et al. "An Oxidative Stress Biomarker for Noise-Associated Stroke: Evidence from Human and Mice." *Noise and Health*, vol. 28, no. 133, 2026, pp. 945-58, https://doi.org/10.4103/nah.nah_239_25.

4. Sang, Yinghao. "Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs." *2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, IEEE, 2025, pp. 417-20, https://doi.org/10.1109/aiea66061.2025.11160531.

5. Lewis, Patrick, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459-74.

6. Karpukhin, Vladimir, et al. "Dense Passage Retrieval for Open-Domain Question Answering." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 6769-81.

7. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. National Academies Press, 2019.

8. Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, 2016, article 160018.

9. Hernán, Miguel A., and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

10. Rothman, Kenneth J., Sander Greenland, and Timothy L. Lash. *Modern Epidemiology*. 3rd ed., Lippincott Williams & Wilkins, 2008.

11. Benjamini, Yoav, and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B*, vol. 57, no. 1, 1995, pp. 289-300.

12. Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. "Machine Learning in Medicine." *New England Journal of Medicine*, vol. 380, 2019, pp. 1347-58.

13. Wiens, Jenna, et al. "Do No Harm: A Roadmap for Responsible Machine Learning for Health Care." *Nature Medicine*, vol. 25, 2019, pp. 1337-40.

14. Rusu, Andrei A., et al. "Policy Distillation." *arXiv*, 2015, arXiv:1511.06295.