



WHITE PAPER

Measuring What Matters: Causaly's 5-Dimensional Framework for Benchmarking Agentic Scientific Outputs

Version 1.0 (8th August, 2025)

Table of Contents

Executive Summary	3
Quantitative findings based on LitQA2 Benchmark	3
Qualitative findings based on Causaly 5D Benchmark	5
Traditional Benchmarks - A useful starting point, not the finish line	7
Causaly's 5Dimension evaluation framework	8
From scoreboard to substance - Applying the framework	9
Key findings	9
Dimension 2 - Depth of Analysis	11
Dimension 3 - Argument Structure	14
Dimension 4 - Assumption Transparency	15
Dimension 5 - Limitation Disclosure	17
Conclusion	19
Appendices	20
Appendix A - Five Dimension Rubric (full scoring tables)	20
Appendix B - sample reports	21

Executive Summary

Large-language-model agents can now [plan research](#), [critique their own reasoning](#), and [draft multi-page scientific reviews in minutes](#), shrinking weeks of effort to hours. But velocity is not valuable unless the output is trustworthy. Single-score benchmarks (accuracy, F1, “Humanity’s Last Exam”) reveal whether a model found the right sentence, but they say little about the [depth of causal reasoning](#), [transparency of assumptions](#), [visibility of evidence gaps](#), or [defensibility in front of peers, boards, and regulators](#). In biomedical R&D, where an AI-generated insight can steer multi-million-dollar programs or appear in regulatory dossiers, those blind spots are not acceptable.

Causaly’s point of view is to judge agentic outputs the same way scientists judge a paper. [Does the argument hold together? Are claims traceable to high-quality sources? Where is the uncertainty, and what experiments would close the gap?](#) To answer those questions, we introduce a [five-dimensional \(5D\) evaluation framework](#) purpose-built for scientific research:

- Factual Accuracy – Are atomic claims complete, verifiable, and properly cited?
- Depth of Analysis – Does the report reason through multi-step biology and weigh conflicting evidence?
- Argument Structure – Is there a coherent chain from framing question to conclusion?
- Assumption Transparency – Are extrapolations surfaced, risk-graded, and linked to next steps?
- Limitation Disclosure – Where does evidence run out, and what data would close the gap?



But velocity is not valuable unless the output is trustworthy.

Traditional benchmarks remain a [necessary first gate](#) - we use LitQA2 to stress-test retrieval and grounding – but they are [not the finish line](#). This whitepaper explains how the 5D-rubric fills those blind spots. It illustrates this with a head-to-head comparison of Causaly Deep Research versus other tools, reporting quantitative gains on LitQA2 alongside qualitative gains in reasoning clarity.

Quantitative findings based on LitQA2 Benchmark

LitQA2 asks 199 biomedical questions whose answers are buried in full-length scientific papers. Systems must retrieve the passage and choose the correct multiple-choice answer, testing retrieval, grounding, and reasoning. This models the real-life use case of searching for rare knowledge among millions of documents that is needed to generate answers for scientific questions. Causaly Deep Research outperforms all tested products and models in this benchmark.

Key Take-Aways from the LitQA2 Benchmark

Our latest Causaly Deep Research v2 (DRv2) decisively outperforms every AI system we tested and closes much of the gap to human reviewers.

Dimension	Human Expert	Causaly DR v2	Best Other AI	Margin vs. Best AI
Coverage	92%	86%	75%	+11 pp
Accuracy	70%	76%	60%	+16 pp
Precision	76%	88%	81%	+7 pp

LitQA2 Benchmark Findings

Highlights

- **Highest overall scores** – Tops the chart on both accuracy and precision while achieving near-human coverage.
 - **Domain depth pays off** – Outperforms general-purpose LLM stacks by wide margins because the pipeline is tuned for biomedical literature retrieval and causal reasoning.
 - **Credible at scale** – Precision that exceeds human reviewers means fewer false positives, critical for regulated R&D settings.
- **Breadth without trade-offs** – Maintains broad evidence coverage (86 %) without sacrificing correctness.
 - **Rapid progress curve** – DRv2 leap-frogs the prior Causaly build (v1), showing that our focused investments compound over successive releases.

On LitQA2, Causaly Deep Research v2 delivers human-grade reliability, setting a new benchmark for scientific deep-research systems.

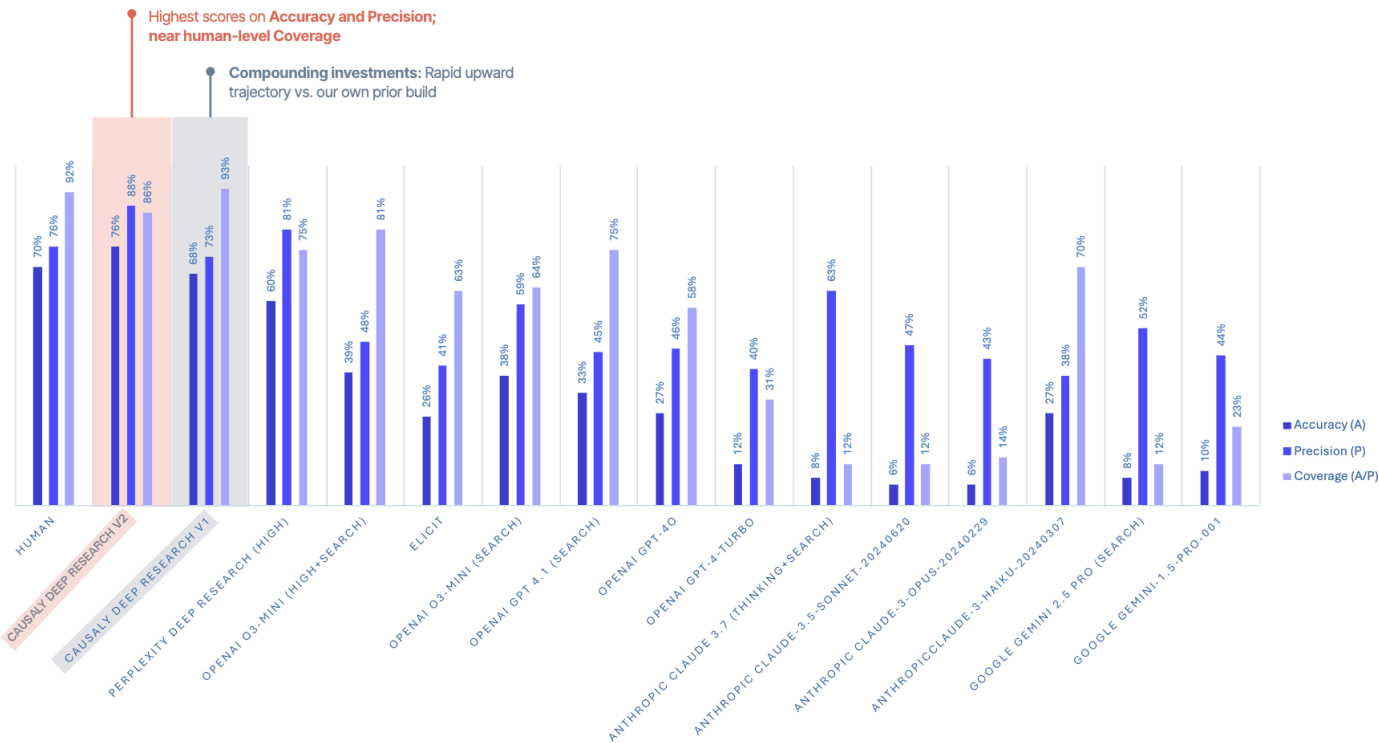


Figure 1: Test of several AI systems against the LitQA2 benchmark. Causaly DRv1, v2, and Perplexity DR (High) benchmarks run by Causaly while remaining benchmarks from <https://arxiv.org/pdf/2407.10362v1> and <https://arxiv.org/pdf/2409.13740>

Qualitative findings based on Causaly 5D Framework

Clearing LitQA2 tells us that an agent can find the right passage and select the right fact. The next question is, can the same agent weave hundreds of such facts into a coherent, defensible scientific argument?

To answer that, we apply our five-dimensional rubric - Factual Accuracy, Depth of Analysis, Argument Structure, Assumption Transparency, and Limitation Disclosure - on a real-world task that scientists routinely face:

“Analyze the involvement and causal biology for TGFB1, GDF5,FGF18, CTSK, IL11, DPEP1, DIABLO, CRHR1, MAPT and TNFSF15 in Osteoarthritis”

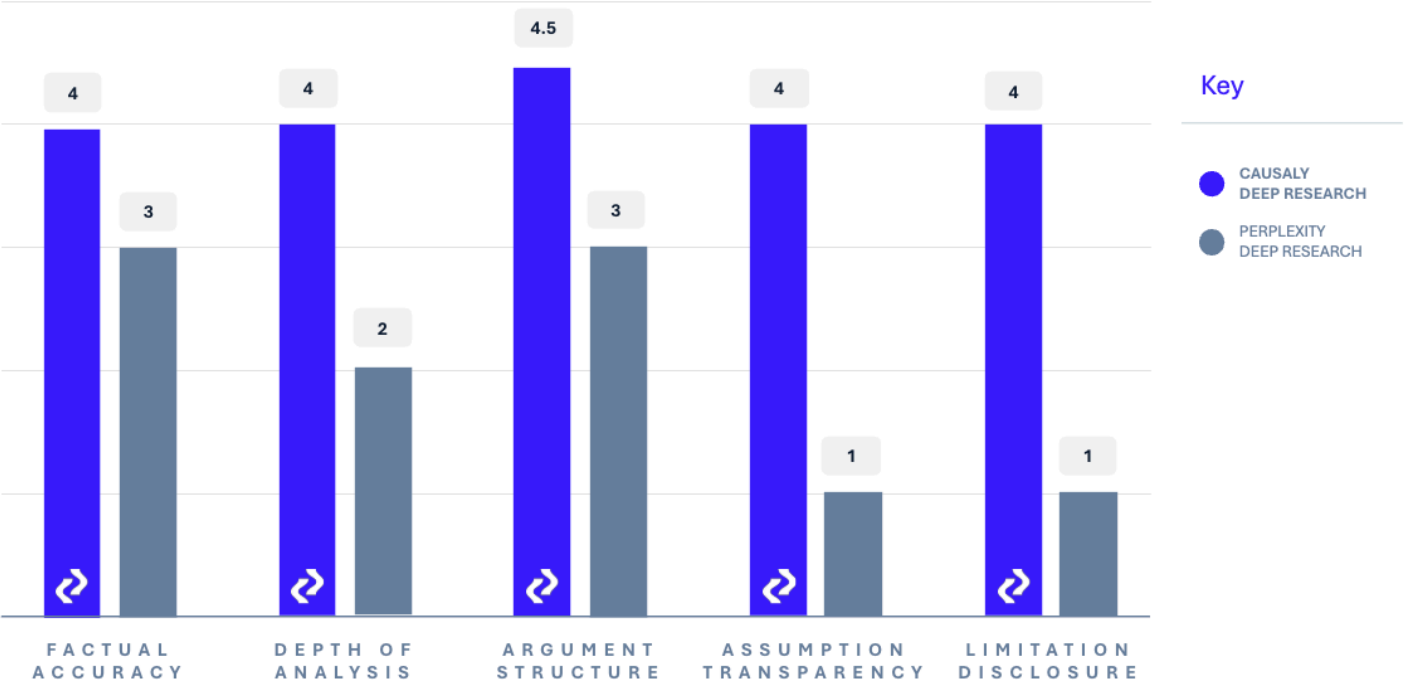


Figure 2: Causaly DRv2 compared against Perplexity Deep Research (high reasoning) on the 5D framework

High-level summary (details in later chapters):

- **Factual Accuracy** – Application of the LitQA2 benchmark falls under this dimension. Causaly cross-checks sampled claim–source pairs (Groundness Table) with no broken links, while Perplexity DR relies on unsystematic inline citations; hence the accuracy gap.
- **Depth of Analysis** – Causaly links genetics → function → mode-of-action (MoA) and explicitly downgrades weak genes after weighing evidence quality. Perplexity DR delivers descriptive pathways but little evidence appraisal.
- **Argument Structure** – Causaly follows a Cochrane-style arc with summary and certainty labels. Perplexity DR offers gene-by-gene prose lacking a unifying summary or clear reasoning chain.
- **Assumption Transparency** – Causaly makes missing data or assumptions explicit in its Limitations table. Perplexity DR leaves assumptions implicit, forcing reviewers to guess.
- **Limitation Disclosure** – Causaly flags gaps and explains impact. Perplexity DR provides no dedicated Limitations section, so gaps stay hidden.

Table: Evidence Status Summary for Each Gene

Gene	Genetic Association & Causality	Functional & Mechanistic Evidence	Mode-of-Action/Pathway	Outstanding Evidence Gaps
TGFB1	Robust GWAS, fine-mapping, eQTL (Tachmazidou Ioanna et al., 2019; Rice Sarah J et al., 2021; Lu Ning et al., 2017)	Strong animal, in vitro, and human studies (Chen Rebecca et al., 2015; Thomas M et al., 2019)	ECM, TGF-β, cartilage and bone homeostasis (Plaas A et al., 2011; Li Jun et al., 2012)	None critical
GDF5	Very strong GWAS (rs143383), fine-mapping, eQTL (Zhang Rui et al., 2015; Wang Yue-peng et al., 2023)	Strong in vivo, LOF, and supplementation data (Parrish W R et al., 2016; Kania Karolina et al., 2020)	Chondrogenesis, cartilage maintenance (Takahata Yoshifumi et al., 2022; Pregizer Steven K et al., 2018)	None critical
FGF18	Moderate GWAS, some eQTL (Tachmazidou Ioanna et al., 2019; Zhao Tianyun et al., 2020)	Strong clinical and preclinical anabolic data (Moore E E et al., 2005; Lohmander L Stefan et al., 2014; Mori Yoshifumi et al., 2014)	Cartilage proliferation, repair (Yao Xudong et al., 2018; Coutinho de Almeida Rodrigo et al., 2020)	None critical
CTSK	Moderate GWAS, MR (Tachmazidou Ioanna et al., 2019; Borge Gábor et al., 2020)	Strong in vivo inhibition/KO, biomarker (CTX II) (Layman et al., 2014)	Cartilage matrix catabolism (Konttinen et al., 2014)	None critical

Figure 3 Example of Causaly DRv2, structured representation of evidence



Traditional benchmarks – a useful starting point, not the finish line

Industry-standard leaderboards such as [LitQA2](#) or [Humanity's Last Exam](#) style multiple-choice sets still play an important role. They let teams:

- **Baseline quickly and objectively** – A single accuracy-precision-coverage scoreboard shows whether a new retrieval stack or model tuning helps or hurts before anyone sinks time into deeper review.
- **Stress-test core skills** – Good question banks, force an agent to locate the right passage inside full-text papers and ground a specific answer, exercising the retrieval-plus-reasoning muscles every scientific report relies on.
- **Debug and track progress reproducibly** – Because the inputs and scoring stay fixed, results are repeatable across time and teams, making it easy to spot regressions or confirm improvements.

At Causaly, we run these tests as a [first-pass gate](#). If an agent cannot clear a high bar on factual retrieval, it is not ready for production. But passing the gate is only the starting point. While static Q&A metrics can tell you whether a model found the right sentence, they cannot tell you whether a 10-page agentic report shows sound causal reasoning, surfaces its assumptions, or reveals where evidence is weak.



If an agent cannot clear a high bar on factual retrieval, it is not ready for production.

Data Notes

LitQA2 is a sub-task of the LAB-Bench suite. It evaluates whether an AI agent can retrieve a specific passage from a full scientific paper and answer a multiple-choice question supported by that passage.

- **Scope & design** – The dataset contains 199 recent, peer-reviewed papers chosen to avoid overlap with LLM training corpora. Each item pairs a question with challenging distractors, so success requires grounded retrieval rather than parametric recall.
- **Why it's hard** – Relevant facts are often buried deep in the PDF, not in the abstract; models must perform fine-grained search and then reason just enough to pick the single defensible answer.
- **Why we rely on it** – LitQA2 simulates the “find-the-evidence, justify-the-claim” micro-task that underpins every longer scientific review, making it an ideal first-gate benchmark for any agentic research system before deeper rubric scoring.

For more information on the benchmark, see the original arXiv papers.¹



Causaly's 5-Dimensional Evaluation Framework

To bridge the gap between quick-fire LLM metrics and the rigor scientists demand, we created a [five-dimensional scoring framework](#). It extends familiar scientific peer-review criteria of evidence, logic, and limitations into a scorecard that any agentic research system must clear before its output can guide real R&D decisions. By grading agentic outputs on these five dimensions, organizations gain a defensible, repeatable standard for deciding when to trust and when to challenge AI-derived insights.

Dimension 1: Factual Accuracy

[Are the individual claims and facts correct and properly cited?](#) Mis-citation creates additional effort to validate claims, and hallucinations erode trust. We check that statements in the report are supported by real studies from reliable sources, not hallucinated or misrepresented. We also look at the quality and traceability of the references used, not just whether there is a citation, but whether it comes from credible clinical, preclinical, or mechanistic evidence. LitQA2 is part of assessing this dimension.

Dimension 2: Depth of Analysis

[Does the report go beyond summarizing facts to actually reason through biology?](#) A strong report doesn't just say "X is linked to osteoarthritis"; it explains how, through what pathway, and under what conditions. We evaluate whether the report reasons through multi-step mechanisms, compares alternative hypotheses, or integrates different types of evidence (e.g., genetics, animal models, patient data). Deeper analysis means stronger scientific value.

Dimension 3: Argument Structure

[Is the report logically organized like a scientific argument or just a list of findings?](#) Good science has structure. We look for a clear framing question, a logical build-up of sub-claims, and a conclusion based on evidence. Strong reports group related ideas, highlight intermediate steps in reasoning, and often use visual elements (like tables or flow diagrams) to guide the reader through complex thinking.

Dimension 4: Assumption Transparency

[Does the report clearly state what is being assumed and how much it matters?](#) Scientific reasoning always involves assumptions: Are we extrapolating from animal studies? Are we assuming a certain trial result applies to all patients? We score reports higher when these assumptions are made explicit, and especially when they're graded by how confident we are in them and how critical they are to the overall conclusion.

Dimension 5: Limitation Disclosure

[Does the report tell you where the evidence runs out?](#) No study is perfect, and no summary is exhaustive. We look for whether a report calls out what's missing, for example, "no human trial data exists" or "this mechanism hasn't been tested in cartilage tissue." Clear disclosure of limitations helps scientists understand what's solid and what still needs further investigation.

Detailed scoring rubrics and examples appear in Appendix A.²

From scoreboard to substance — Applying the framework

Passing LitQA2 tells us that an agent can find the right passage and select the right fact. The next question is, *can the same agent weave hundreds of such facts into a coherent, defensible scientific argument?*

To answer that, *we apply our five-dimensional rubric* — Factual Accuracy, Depth of Analysis, Argument Structure, Assumption Transparency, and Limitation Disclosure — *on a real-world task that scientists routinely face.*

"Analyze the involvement and causal biology for TGFB1, GDF5, FGF18, CTSK, IL11, DPEP1, DIABLO, CRHR1, MAPT and TNFSF15 in Osteoarthritis"

Why This Example?

- **Breadth & complexity** – Ten genes with evidence spanning genetics, animal models, mechanistic biology, and early clinical signals provide an ideal stress test for an agent's ability to connect multiple evidence layers in one coherent argument.
- **High stakes** – Target assessment guides go/no-go decisions worth millions in pre-clinical budgets.
- **Comparable outputs** – We generated full reports from Causaly Deep Research and Perplexity Deep Research, then scored each against the rubric

Data Notes and Methodology

This evaluation compares Causaly Deep Research v2 (July 2025 build) and Perplexity Deep Research, using the osteoarthritis question detailed earlier as the prompt. Both systems were tested on the same day, 18 July 2025, to ensure consistency.

Each output was scored against the five evaluation dimensions using the O3 model from OpenAI. To maintain impartiality, the assessment was conducted in a 'blind' fashion, with no indication of which system generated which report. An independent human assessment of the same reports will follow.



Key Findings

Dimension (0-5)	Causaly Deep Research	Perplexity Deep Research	Why it earned the score (key excerpts)
1. Factual Accuracy	4	3	Causaly includes a Groundness Table that cross-checks sampled claim-source pairs, a hallmark of auditability. No broken links found. Minor residual risk that not every claim is sampled prevents a "5". Perplexity Deep Research cite primary literature inline (e.g. TGFB1 pathway, rs75621460 SNP) but offer no systematic source audit; reviewers must manually verify.
2. Depth of Analysis	4	2	Causaly weaves genetics → functional → MoA and explicitly downgrades weak genes (DIABLO, MAPT, TNFSF15) after weighing missing fine-mapping & perturbation evidence. That evidence-quality appraisal meets rubric "4"; lack of formal risk-of-bias scoring keeps it from "5". Perplexity Deep Research list mechanistic color (e.g., TGF-β ALK1/ALK5 switch) but rarely contrast evidence strength or address contradictions; it reads like a literature digest (example bullets under "Integrative Pathogenic Mechanisms").
3. Argument Structure	4.5	3	Causaly follows a Cochrane-style arc: Direct Answer → Evidence Status Table → per-gene triad → Limitations Table → concise conclusion listing "validated" vs "tentative" targets. Logical flow and sign-posted certainty. Perplexity Deep Research open with an "Overview" and head straight into mixed-length gene paragraphs; no summary table or certainty labels (see section heading block). The reasoning chain is only partially visible.
4. Assumption Transparency	4	1	Causaly surfaces missing evidence as explicit assumptions/gaps (e.g., ✗ fine-mapping for DIABLO) inside the Limitations Table with impact on interpretation. Perplexity Deep Research never label assumptions; the reviewer must infer them from the prose. No dedicated table.
5. Limitation Disclosure	4	1	Causaly tags every evidence dimension ✓/⚠/✗ and explains scientific impact (e.g., temporal-biomarker void for TNFSF15). Actionable but not yet linked to follow-up experiments. Perplexity Deep Research have no "Limitations" section; gaps like lack of functional validation are never called out.

Dimension 2 – Depth of Analysis

Reviewer verdict: Causaly links genetics, functional data, and mechanism in multi-step chains and downgrades weak evidence, though it stops short of a formal bias appraisal. Perplexity Deep Research is largely descriptive; mechanistic details are present, but evidence strength and contradictions are not evaluated.

- Weigh evidence quality (in-vivo > in-vitro > in-silico).
- Surface contradictions and residual uncertainty.

The rubric, therefore, awards higher scores when the report not only lists studies but integrates them into a logical, quality-aware narrative (Score 4) and, at the top end, reconciles counter-evidence and performs a nuanced appraisal (Score 5).

Drug-discovery go/no-go calls hinge on causal rather than correlative evidence. A robust synthesis must:

- Stitch multi-hop chains (e.g., GWAS → perturbation → phenotype).

Table: Evidence Status Summary for Each Gene

Gene	Genetic Association & Causality	Functional & Mechanistic Evidence	Mode-of-Action/Pathway	Outstanding Evidence Gaps
TGFB1	Robust GWAS, fine-mapping, eQTL (Tachmazidou Ioanna et al., 2019; Rice Sarah J et al., 2021; Lu Ning et al., 2017)	Strong animal, in vitro, and human studies (Chen Rebecca et al., 2015; Thomas M et al., 2019)	ECM, TGF-β, cartilage and bone homeostasis (Plaas A et al., 2011; Li Jun et al., 2012)	None critical
GDF5	Very strong GWAS (rs143383), fine-mapping, eQTL (Zhang Rui et al., 2015; Wang Yue-peng et al., 2023)	Strong in vivo, LOF, and supplementation data (Parrish W R et al., 2016; Kania Karolina et al., 2020)	Chondrogenesis, cartilage maintenance (Takahata Yoshifumi et al., 2022; Pregizer Steven K et al., 2018)	None critical
FGF18	Moderate GWAS, some eQTL (Tachmazidou Ioanna et al., 2019; Zhao Tianyun et al., 2020)	Strong clinical and preclinical anabolic data (Moore E E et al., 2005; Lohmander L Stefan et al., 2014; Mori Yoshifumi et al., 2014)	Cartilage proliferation, repair (Yao Xudong et al., 2018; Coutinho de Almeida Rodrigo et al., 2020)	None critical
CTSK	Moderate GWAS, MR (Tachmazidou Ioanna et al., 2019; Borge Borge et al., 2020)	Strong in vivo inhibition/KO, biomarker (CTV II) (Hayami et al., 2019)	Cartilage matrix catabolism (Kortgen et al., 2019)	None critical

Figure 3 Example of Causaly DRv2, structured representation of evidence



Scores & Justification

System	Depth score	Rationale
Causaly	4	Satisfies all Score 4 criteria: multi-hop chains, explicit evidence gaps, qualitative certainty labels, and discussion of contradictory data (e.g., context-dependent IL-11). Falls short of "5" because it does not yet run a formal risk-of-bias framework (e.g., GRADE) across studies.
Perplexity DR	2	Provides mechanistic color but largely descriptive: no weighting of study design, sparse treatment of contradictions, and absence of gap table. Meets rubric "2" (some synthesis) but not "3" because causal links and quality discussion are inconsistent.

Scoring Rubric Recap

- 0 – Pure extraction; no interpretation.
- 1 – Facts regrouped but not related.
- 2 – Some synthesis (grouping or chronology).
- 3 – Causal reasoning chain shown ($X \rightarrow Y \rightarrow Z$).
- 4 – Weighs evidence quality; handles uncertainty/contradictions.
- 5 – Full expert-level synthesis: counter-arguments, subgroup analysis, nuanced appraisal.



Deep dive comparison

Sub-criterion	Causaly Deep Research	Perplexity Deep Research	Impact on scientific decision
Multi-hop causal chains	Per-gene triad "Genetic Causality → Functional Biology → Mode-of-Action" is explicit; see TGFB1 block where GWAS, dose-dependent SMAD vs non-SMAD signaling, and tissue response are chained in order.	Mechanistic prose exists (e.g., TGFB1 ALK1/ALK5 switch) but presented in narrative form without tying back to genetic origin or functional validation.	Causaly let reviewers trace cause→ mechanism→ effect. Perplexity Deep Research forces manual reconstruction.
Evidence-quality appraisal	Limitations table grades which evidence dimensions are met or missing (e.g., ✗ functional perturbation for DIABLO, MAPT, TNFSF15).	No systematic appraisal; all cited studies are treated equally regardless of sample size or model quality.	Causaly prevents over-weighting weak loci. Perplexity Deep Research may inflate speculative targets.
Handling weak / contradictory evidence	Explicit downgrades: "DIABLO has mechanistic plausibility but lacks fine-mapping/colocalization—remain speculative".	Mentions limited evidence (e.g., "direct evidence for MAPT involvement is limited") but still narrates mechanistic relevance without adjusting confidence.	Causaly flags risk before pipeline progression; Perplexity Deep Research may pass borderline genes.
Breadth across domains	Integrates human GWAS, eQTL, KO-mouse, biomarker cohorts and pathway data (see IL11 synthesis with genetic + signaling context lines).	Covers diverse pathways but rarely links them to corresponding genetic or animal evidence; domain connections are implicit.	Broader, linked evidence raises the likelihood that in-vivo validation will succeed.
Uncertainty language & gradation	Per-gene Certainty: High / Moderate / Low labels help prioritize follow-ups.	No explicit certainty scale; reviewer must infer from tone.	Causaly short-circuits debate on which gene to resource first.

Dimension 3 – Argument Structure

Reviewer verdict: Causaly reads like a mini-systematic review. I can trace conclusions from the framing question to the evidence table to the limitations. Perplexity Deep Research offers useful prose, but the logical scaffold and certainty grading are missing, so I would have to reconstruct the argument myself.

Portfolio decisions depend on quickly grasping how the evidence supports (or refutes) a claim. A clear argumentative scaffold, question → sub-claims → evidence → weighed conclusion, reduces reviewer time, exposes weak links early, and boosts confidence when recommending pipeline advancement.

Scores & Justification

System	Score (0-5)	Rationale
Causaly DR	4.5	Meets all rubric "4" indicators—clear framing, grouped sub-claims, evidence tables, linking phrases—and partially fulfils "5" by downgrading weak genes and acknowledging counter-evidence. Lacks a formal counter-argument weighing table to reach a 5.
Perplexity DR	3.0	Achieves rubric "3": sub-claims exist and some grouping is visible, but framing is implicit, no certainty or limitations tables, and logical flow must be inferred.

Scoring Rubric Recap

0 – Disorganized prose; no framing.
1 – Question only implied.
2 – Question stated; flat evidence list.
3 – Sub-claims grouped; partial logic visible.
4 – Clear progression claim → sub-claim → evidence with linking phrases.
5 – Full scientific arc (intro, hypotheses, grouped evidence, counter-points, conclusion).

Deep dive comparison

Sub-criterion	Causaly Deep Research	Perplexity Deep Research	Impact on scientific decision
Counter-evidence handling	Downgrades DIABLO/MAPT/TNFSF15 as speculative with rationale	Mentions weak evidence but keeps narrative weight equal	Prevents over-interpretation of weak targets.
Framing question / hypothesis	Opens with explicit scope & uncertainty in Direct Answer	Hypothesis never stated; pathways listed generically	Faster reviewer orientation.
Top-level scaffold	Evidence Status table (genes × evidence dimensions) gives instant map	Continuous prose; uneven headings	Tabular scaffold short-circuits mental parsing.
Logical progression	Per-gene triad links genetic → functional → MoA; certainty label closes loop	Mixed narrative; causal links implicit	Traceable reasoning vs. inference burden.
Linking phrases & sign-posts	Frequent “therefore”, “this suggests”; icons ✓/⚠/✗ guide flow	Rare transitions; no visual cues	Visual sign-posts speed comprehension.

Dimension 4 – Assumption Transparency

Reviewer verdict: Causaly surfaces key assumptions, tags their impact, and suggests follow-up experiments; this is sufficient for risk assessment. Perplexity Deep Research does not expose their assumptions, leaving potential logical gaps hidden.

Hidden assumptions such as species translatability, in-silico model reliability, and untested causal direction can derail downstream validation. A high-stakes go/no-go demands that these suppositions are explicit, linked to claims, and risk graded.

Assumptions Table

Assumptions Statement	Role in Answer Construction	Criticality	Confidence Level	Scientific veracity / Gap Description
MAPT is not a causal OA gene but may be a biomarker or involved in OA-related neurodegeneration	Inclusion as a biomarker, not a causal gene	Minor	Confident	No GWAS or functional causality
TNFSF15 is functionally involved in OA but not genetically causal	Supports inclusion as OA effector gene	Important	Confident	Transcriptomic and functional evidence only
DIABLO is not involved in OA	Excludes DIABLO from OA gene list	Minor	Highly confident	No evidence found

Figure 5 Example of Assumptions Table from Causaly Deep Research v2.

Scores & Justification

System	Score (0-5)	Rationale
Causaly DR	4	Dedicated table lists assumptions with impact/confidence tags (High / Weak) and proposed resolution steps. Each assumption references the claim it underpins ("A2 supports Functional Biology section")—fulfils rubric "Context linkage". Falls short of "5" because follow-up loop is described but not executed in the report
Perplexity DR	1	No assumption table; extrapolations (e.g., neuro-MAPT relevance to cartilage) are embedded in prose with no risk tags. Reviewers must infer critical leaps—meets only rubric score "1" (occasional implicit mention).

Scoring Rubric Recap

- 0 – No assumptions surfaced
- 1 – Occasional implicit assumption called out ad-hoc
- 2 – List of assumptions but no impact/confidence grading
- 3 – Explicit table with impact or confidence tags
- 4 – Table links each assumption to the claim it supports
- 5 – As "4" + resolution loop (proposed actions to retire or validate assumptions)

Deep dive comparison

Sub-criterion	Causaly Deep Research	Perplexity Deep Research	Impact on scientific decision
Explicit surfacing	3-row assumption table visible	None	Transparency vs. reviewer guesswork
Risk/Confidence grading	High/Med/Low × Strong/Moderate/Weak tags	Absent	Prioritizes lab follow-ups
Resolution plan	Proposed in-silico or wet-lab step for each high-impact gap	None	Turns speculation into work plan

Dimension 5 – Limitation Disclosure

Reviewer verdict: Causaly maps evidence gaps gene-by-gene and proposes concrete next steps, giving a clear agenda for further work. Perplexity Deep Research offers no structured gap analysis, so blind spots remain undisclosed.

Every scientific question implicitly defines a **blueprint of evidence dimensions**, a checklist of data slices that must be inspected before any conclusion is trustworthy (e.g., genetic support, functional assays, structural data, translational markers, etc.).

The Limitations Table exists to expose where that checklist is still incomplete after a Deep Research run.

- **Purpose** – Turns a free-text answer into an auditable matrix that flags each required evidence dimension as *Found*, *Weak*, or *Missing*.

- **Why it matters** – Scientists can immediately see whether the report skipped a critical line of evidence (say, epitope mapping or cross-species replication).
- **What's evaluated** – For every dimension on the blueprint, we check:
 - a. Presence – Was any supporting data retrieved?
 - b. Strength – Is the data conclusive or merely suggestive?
 - c. Next step – If missing, what computational action could close the gap (e.g., targeted PubMed sweep, structure-model docking)?

- **Difference from the Assumptions Table** (previous chapter) – Limitations focus on **data availability** (did we fetch it?); Assumptions focus on **reasoning logic** (how we stitched the fetched data together).
- **Benchmark angle** – We score systems on how transparently and completely they surface these gaps. Exposing these gaps is essential for the scientific process.

Limitations Table		
Evidence Dimension	Addressed Status	Notes
GWAS association/fine-mapping/colocalization	✓ for TGFB1, GDF5, FGF18, CTSK, IL11, DPEP1, CRHR1	✗ for DIABLO, MAPT, TNFSF15
Functional/animal model perturbation	✓ for TGFB1, GDF5, FGF18, CTSK, IL11, CRHR1	✗ for DPEP1, DIABLO, MAPT, TNFSF15
Temporal-sequence (biomarker cohort)	✓ for CTX-II (CTSK)	✗ for IL-11, TNFSF15, DIABLO, MAPT
Human phenotypic convergence	⚠ partial (GDF5, IL11 height)	Not for others
Alternative mechanism exclusion	✗	Not systematically addressed for any gene

Figure 6 Example of Limitations Disclosure table from Causaly Deep Research v2

Scores & Justification

System	Score (0-5)	Rationale
Causaly DR	4	Limitations Table flags each evidence dimension per gene as ✓/⚠/✗ and narrates scientific impact. For every ✗ or high-impact ⚠ cell, a concrete next step is proposed (e.g., fine-mapping study). The report cannot show iterative closure, so not a "5".
Perplexity DR	1	No "Limitations" section; gaps such as missing perturbation data for DIABLO are not flagged. Occasional phrases like "evidence is limited" count only as generic acknowledgement → rubric "1".

Scoring Rubric Recap

0 – No mention of gaps
1 – Generic “every study has limits” paragraph
2 – Lists some gaps but not mapped to evidence
3 – Dedicated Limitations section; gaps are flagged
4 – Actionable next-step suggestions for each high-impact gap
5 – Iterative closure loop: system tracks whether gaps are filled in subsequent cycles

Expert comparison

Sub-criterion	Causaly Deep Research	Perplexity Deep Research	Impact on scientific decision
Dimension-level gap flagging	Schema table with icons for each missing evidence type	None	Instant clarity on where to invest
Actionable next steps	“Launch Seurat deconvolution...”, “Search LINCS longitudinal datasets” for each ✗ gap	Not provided	Converts gaps to task list
Narrative boundary setting	States population/species limits (e.g., no human cohort)	Absent	Prevents over-generalization
Closure behavior	Described but static in report (future agent loop)	Not present	Drives continuous improvement
Dimension-level gap flagging	Schema table with icons for each missing evidence type	None	Instant clarity on where to invest

Conclusion

Relying on single-number accuracy scores when selecting an AI research partner is no longer sufficient. A five-dimensional evaluation exposes substantive differences in reasoning quality, transparency, and scientific usefulness.

Across both quantitative (LitQA2) and qualitative (osteoarthritis deep-dive) evidence, [Causaly Deep Research](#) demonstrates leading performance, pairing high factual correctness with the structured argumentation and disclosure practices required for defensible R&D decisions.



Appendices

¹ More information on the benchmark can be found here: <https://arxiv.org/pdf/2407.10362v1>,
<https://arxiv.org/pdf/2409.13740>

² Appendix A—Five Dimension Rubric (full scoring tables)

Dimension	Definition	How to Evaluate	Scoring rubric
Factual Accuracy	Are individual claims verifiable and correctly grounded in literature?	- Manual annotation by scientific curators or SMEs - Use external QA benchmarks (e.g., BioASQ, LitQA) - Spot-check claims with inline citations	0 – Many hallucinations, no citations 1 – Scattered citations, error rate > 30 % 2 – Mostly correct but unsupported claims remain 3 – ≥ 90 % atomic facts correct; minor citation gaps 4 – All claims verifiable; high-quality primary sources; zero broken links 5 – As “4” plus audit artefacts (e.g., Groundness table sampling claim–source pairs) and explicit hallucination guardrails
Depth of Analysis	Does the report reason through biology, not just state it?	- Score based on depth and structure of causal chains - Look for comparative logic (e.g., “A is more likely than B due to Z”) - Visuals showing chains	0 – Pure extraction; no interpretation 1 – Facts regrouped but not related 2 – Some synthesis (grouping or chronology) 3 – Causal reasoning chain shown ($X \rightarrow Y \rightarrow Z$) 4 – Weighs evidence quality; handles uncertainty/contradictions 5 – Full expert-level synthesis: counter-arguments, subgroup analysis, nuanced appraisal
Argument Structure	Is the report built like a scientific argument, not an article summary?	- Evaluate presence of logical flow: intro → reasoning → evidence → conclusion - Higher score if visual diagrams support argument progression	0 – Disorganized prose; no framing 1 – Question only implied 2 – Question stated; flat evidence list 3 – Sub-claims grouped; partial logic visible 4 – Clear progression claim → sub-claim → evidence with linking phrases 5 – Full scientific arc (intro, hypotheses, grouped evidence, counter-points, conclusion)
Assumption Transparency	Does the system state what it assumes, and how critical those assumptions are?	- Check if assumptions are surfaced clearly and tied to claims - Higher score if visual layout supports reviewer understanding of assumption role	0 – No assumptions surfaced 1 – Occasional implicit assumption called out ad-hoc 2 – List of assumptions but no impact/confidence grading 3 – Explicit table with impact or confidence tags 4 – Table links each assumption to the claim it supports 5 – As “4” + resolution loop (proposed actions to retire or validate assumptions)

Appendices

Dimension	Definition	How to Evaluate	Scoring rubric
Limitation Disclosure	Are the boundaries of evidence made clear to the reader?	- Look for "Limitations" section or inline disclaimers - Assess alignment between known gaps and what the system flags - Reward use of limitation tables or maps	0 – No mention of gaps 1 – Generic "every study has limits" paragraph 2 – Lists some gaps but not mapped to evidence 3 – Dedicated Limitations section; gaps flagged ✓/⚠/✗ per evidence dimension 4 – Actionable next-step suggestions for each high-impact gap 5 – Iterative closure loop: system tracks whether gaps are filled in subsequent cycles

Appendix B: Sample reports

The Deep Research reports that were used for the evaluation are available on request.