Syno: A Citation-Faithful Literature Review Agent for Biomedical Research

Syntactiq

June 2026

Abstract

The first generation of AI literature-review tools has miscounted its own failure mode. The field has trained itself to watch for invented papers; the harder problem is that real papers are mis-described. The citation resolves and the DOI checks out, yet the sentence the AI built around it overstates, mis-populates, or reverses what the paper actually reports. Peer-reviewed audits now document this as the dominant failure pattern in biomedical work, not citation fabrication.

On a pre-registered, diabetes-scoped external benchmark (ten systematic-review questions, graded by an independent three-model panel of Anthropic Claude and Google Gemini), Syno led the strongest purpose-built peer (Elicit) on per-claim numeric fidelity by roughly **+19 points on this slice**, a lead that held in 9 of 10 questions and across every judge subset. We read the *relative, judge-robust* lead as the signal rather than any single absolute number. Rates move with judge strictness (Syno 92% on the lenient majority panel, 69% on the strict single-judge read), bracketing the careful-human quotation-accuracy band (~75–83%) [13][14][15], which we cite only as loose context for scale, not as a like-for-like equivalence. This is one corroborating data point rather than the thesis: scope (one disease, one run, one capture per tool), the judge-calibration spread, and each tool's differing evidence base are disclosed in §7 and Annex A. Syno also runs slower, and likely costs more per report, than single-pass tools, a deliberate accuracy-first trade. Anyone can re-grade from the published questions and raw verdicts.

Syno is a biomedical literature-review agent built around an alternative posture. Every claim is grounded in retrieved passages from a curated biomedical corpus, judged against each cited paper along several independent axes (population, numerics, outcome, recency), and aggregated with study-design–weighted scoring that preserves dissent rather than averaging it away. When the first-pass judgment is unclear or contested, the system escalates to a deeper-reasoning model across model families rather than discarding the signal. Every verdict carries a passage anchor and a policy-version stamp, so a reader can trace any sentence back to the paragraph that justifies it under a known and dated scoring regime.

The limits are explicit. Syno does not yet emit GRADE Summary-of-Findings tables, formal risk-of-bias plots, or MeSH-controlled-vocabulary search artefacts. Those are on the roadmap (§8). What is built today is a defensible foundation: a retrieval-first, claim-level, audit-trailed approach that treats citation faithfulness as a property to be measured rather than asserted.

The goal. Syno optimises for reviews whose every claim a human can verify in seconds rather than hours, not for raw speed over a human reviewer.

Keywords—biomedical literature review, citation faithfulness, evidence synthesis, retrieval-augmented generation, AI agents, PRISMA, audit trail.

1. The Citation Crisis in AI Literature Reviews

1.1 Three peer-reviewed failure-rate findings

Independent evaluations in the past eighteen months have established the empirical floor for off-the-shelf "deep research" agents on biomedical work.

JMIR 2026. Wong, Ong, Merle and Keane evaluated generalist deep-research products on biomedical question-answering. They report that roughly 47–50% of references from Gemini Deep Research and Perplexity AI on medical prompts had fabricated authors or titles, while OpenAI Deep Research achieved ~95% identifiable citations and ~70% completely correct. Over 50% of OpenAI Deep Research's cited statements still contained at least one subtle inaccuracy or misrepresentation of the source [1].

Tow Center for Digital Journalism. A non-medical but methodologically similar audit of eight AI search tools found that more than 60% of queries returned incorrect citations overall, with Grok 3 incorrect roughly 94% of the time and Gemini also performing poorly [2]. This pattern of confident assertion paired with non-resolvable citations is not domain-specific.

Elicit evaluations. Lagisz et al. (2025) and Hilkenmeier et al. (2025) evaluated Elicit's data-extraction outputs against human-graded gold standards across environmental and life-sciences systematic reviews [3]. Interpretation errors (citation drift, scope overreach, missed content) dominated the error mix, rather than citation-existence failures. Elicit was reading real papers; it simply misread them. Lau and Golder's 2025 case-study comparison of Elicit against traditional librarian searching corroborates the pattern at production-relevant scale: average sensitivity around 39.5% against ~94.5% in the original reviews [10]. Bianchi et al.'s 2025 trial of Elicit against human reviewers on RCT data extraction shows narrower but still consistent shortfalls [11].

Human baseline (for context). AI errors should be read against what *careful humans* produce, not against zero. The peer-reviewed quotation- and citation-error literature is sobering: Baethge & Jergas's 2025 meta-analysis (N = 32,074 quotations across 46 studies) finds an overall quotation-error rate of ~16.9%, with ~8.0% classified as *major*, and no improvement over four decades [13]. Jergas & Baethge's 2015 systematic review reports a per-quotation total error rate of ~25.4% (~11.9% major) [14]. Mogull's 2017 single-journal audit lands at ~14.5% per quotation [15]. This does not make AI "as good as humans"; it shows that even the careful baseline an AI tool is being asked to match has only a ~75–83% per-claim accuracy ceiling. We treat that band as a rough *scale reference* for where a tool sits, a different measurement from an AI tool's audited claim fidelity rather than a like-for-like equivalence.

1.2 Why "misinterpretation > fabrication" is the framing the field needs

The Elicit finding generalises. The dominant risk in clinical literature review is a real paper that is **mis-described**, not an invented one. The cited NEJM article exists, the authors are right, and the DOI resolves. The summarising sentence in the AI's prose, however:

- describes the population as "adults with type 2 diabetes" when the trial only enrolled women aged 65 and older;
- reports the primary outcome's hazard ratio while omitting that the confidence interval crosses one;
- conflates a surrogate endpoint (LDL-C reduction) with a clinical endpoint (cardiovascular mortality);
- attributes a 2014 result to "current consensus" despite a 2023 meta-analysis having superseded it;
- or pulls a number from a paper that the target study merely cited, rather than from the target study itself, the citation-drift pattern that surfaces repeatedly in structured-extraction audits.

Each of these failures is locally plausible. Each leaves the citation **valid** in the narrow sense that a librarian's verification (open the PDF, confirm the authors, confirm the title) would clear it. None can be caught without reading the paper carefully against the claim. This is precisely the work the AI was meant to do.

Practitioner discussion in the biomedical literature-review community, including Aaron Tay's ongoing series on AI-assisted literature review, converges on the same point: AI literature-review tools are competent at the parts of the work that resemble summarisation of an explicit abstract sentence, and unreliable at the parts that require reading the methods and results carefully, which is the part that matters for evidence synthesis.

1.3 Adjacent failure modes that compound the problem

Beyond misinterpretation, the literature documents a constellation of related failures, and our own internal calibration work has surfaced a second set of patterns that the published literature has under-named. Syno's design is shaped in direct response to both.

Field-documented failure modes:

1. **Vote-counting consensus.** Tools that surface a simple support/contradict meter risk treating a 50-patient pilot and a 5,000-patient RCT as equal "votes", a practice evidence-synthesis methodologists have spent decades warning against [5].
2. **Cherry-picking.** Where evidence cuts both ways, the easier behaviour for a generative model is to cite the supportive paper and stay quiet about the rest, a pattern repeatedly raised in practitioner reviews of consumer AI research tools.
3. **Effect-size and CI loss.** Even when an abstract is summarised correctly, the methods and results, where effect sizes, CIs, and subgroup nuances live, are routinely garbled.
4. **Trial-name confabulation.** Asked for specific values from named trials (KEYNOTE-006 ORR, CHECKMATE-067 OS), several tools have been documented producing different wrong numbers on identical prompts.
5. **Non-determinism.** The same query, run minutes apart, produces different paper rankings. This is fatal for PRISMA-style reproducibility [4].
6. **Paywall and niche invisibility.** Tools dependent on Semantic Scholar or open-web search miss large fractions of clinically relevant literature: regulatory documents, conference abstracts, older landmark trials behind paywalls.

Under-documented structural failures (named here):

7. **Right number, wrong context.** A class of failure under-named in the field is *attribution drift within a paper*. The cited number exists and the table is real, but the number sits in a row the claim does not invoke. A subgroup figure restated about the overall population; a control-arm value attributed to the treatment arm; a comparator's effect size attributed to the index intervention; a result the paper quoted from another study, restated as the paper's own. A naive "is the value in the paper?" check passes while the verdict is structurally false. This, not citation existence, is the dominant correctness surface in biomedical literature review.
8. **Ground-then-embellish.** A sentence whose core fact is grounded in the cited paper, but where the writing has bolted on an unsupported qualifier: a co-named mechanism the paper does not invoke, a comparator the paper does not assert, a quantifier the paper does not license, or a population-narrowing word the paper does not use. The citation resolves, the underlying claim survives a permissive entail/contradict check, and the embellishment slips through. A single overall verdict on the sentence misses this.
9. **Compound-claim partial match.** Long sentences that stack three to five facts under a single citation are a routine writing pattern in literature review, and a routine source of partial failure: some sub-clauses are grounded in the cited paper, others are fabricated or pulled from a different paper that the retrieval surfaced but the writer did not cite. A

holistic sentence-level entailment judgment cannot distinguish the grounded sub-clauses from the un-grounded ones; an atomic decomposition can.

10. **Population overreach.** T2D evidence restated about T1D; general-adult findings stated about narrow subgroups; intervention results stated about populations the trial excluded. Current AI literature tools, which judge entailment at the document level and rarely surface population as a distinct dimension, typically miss this until a clinician spot-reads the cited methods.

1.4 Where others are advancing the state of the art

The field is not monolithic. **OpenEvidence** has demonstrated the appetite for peer-reviewed-only citation grounding at clinician scale; **Elicit** has set a high bar for structured RCT data extraction [10][11]; **Consensus** has shipped useful research-question summarisation; **Scite** has shown that per-citation classification is a viable layer (Syno operates one step deeper, at per-axis judgment); **Causaly** has built agentic knowledge-graph deep research for pharma medical affairs; **Undermind** has made a serious case on retrieval recall.

Syno's contribution is orthogonal. We are not competing on point-of-care answer latency, on extraction-precision benchmarks, on consumer summarisation, or on retrieval-recall frontiers; those are real and valuable directions, and a mature evidence-synthesis stack will eventually need all of them. The property Syno targets is per-claim auditable trust as a *structural* property of the output, not as a UI affordance.

2. Design Principles

Syno is organised around six principles. Each is a direct response to a documented failure mode and is grounded in established evidence-synthesis methodology (PRISMA 2020 [4], the Cochrane Handbook [5], GRADE [6], SANRA [7]), together with the recent AI-era frameworks: the 2025 Cochrane–Campbell–JBI–CEE Joint Position on AI in evidence synthesis [8] and the PRISMA-trAIce reporting checklist [9].

Principle	Failure mode addressed
Retrieval-first grounding (§2.1)	Citation fabrication
Per-axis judgment (§2.2)	Misinterpretation; population / numerics / outcome / recency slips
Quality-weighted aggregation (§2.3)	Vote-counting; over-weighting small studies
Cross-paper dissent check (§2.4)	Selective citation; suppressed dissent
Multi-tier, cross-family escalation (§2.5)	Single-model confidence collapse; vendor monoculture
Auditable, version-stamped verdicts (§2.6)	Reproducibility failure; undated scoring logic

2.1 Retrieval-first grounding

No claim enters the output without a retrieved passage. The system is structurally prevented from generating prose that is not anchored in a paper the retrieval pipeline has actually fetched, indexed, and surfaced, a stronger constraint than "the model has access to retrieval tools." Planning, drafting, and validation each consume the same fixed evidence set of retrieved papers, identified by a recorded cryptographic fingerprint; the writer cannot fabricate a citation because it cannot name a paper that is not in that set. This structurally prevents out-of-corpus citation fabrication rather than relying on a post-hoc check. The remaining principles handle the harder cases.

2.2 Per-axis judgment

A single entail / neutral / contradict verdict is too coarse to catch misinterpretation. Syno's validator judges each (claim, paper) pair along independent axes that correspond to the failure modes documented in §1:

- **Population.** Does the population the claim references (e.g. "adults with T2D") match the population the paper studied (e.g. women aged ≥65)? Population is treated as a multi-state axis, including the substantively different cases of claim narrower than paper and claim broader than paper, rather than a binary entail/contradict. A paper can entail the broad direction of a claim and still fail the population check.
- **Numerics.** When a claim contains effect sizes, hazard ratios, or percentages, do they align with what the paper reports? The numerics axis distinguishes outcomes including aligned values, direction-only agreement, null-result conflicts (the paper's CI crosses null while the claim asserts an effect), reversal, and the two structurally different forms of missing data, because the downstream synthesis behaviour differs for each.
- **Outcome.** Did the paper measure the outcome the claim names, or a surrogate? Regulator-validated surrogates (HbA1c for glycaemic control) are accepted as direct; LDL-C standing in for cardiovascular mortality is not.
- **Recency.** When a claim invokes "current consensus," is the cited paper actually recent enough, or has it been superseded?

Each axis flags independently. A paper can entail the broad direction of a claim while failing the population axis. In a single-verdict system this paper would silently support the claim. In Syno, the mismatch is preserved through aggregation and surfaced to the writer as a revision hint.

2.3 Quality-weighted aggregation

The Cochrane Handbook is explicit that meta-analytic synthesis must weight evidence by study design and quality, not vote-count [5]. Syno's aggregation follows the established evidence-pyramid ordering: meta-analyses and randomised trials carry the most weight, cohort studies and clinical guidelines next, then case-control and case-series work, with mechanistic in-vitro and animal studies and narrative reviews weighted least. Weights are bounded so no single study can overwhelm a contradicting body of evidence, and the denominator includes every paper the validator examined, so a thin supportive corpus dilutes against the broader retrieved set rather than carrying a claim alone.

A recency factor applies on top in tiered bands. Recent literature carries full weight; mid-era literature is partially down-weighted; older literature is down-weighted further. Foundational landmark studies lose at most one tier of weight and are never silenced. In slower-moving fields the classical evidence remains load-bearing, while in fast-moving fields (newer-generation incretins, SGLT2-CVO outcomes, immune-checkpoint combinations) a recent meta-analysis can override earlier-era results.

These factors are centralised in a single versioned policy. The policy version stamped onto each verdict (§2.6) bumps whenever any weight or band changes, and old-version verdicts are re-judged on the next run rather than silently reused.

2.4 Cross-paper dissent check

Even with grounded retrieval and per-axis judgment, a writer can still cite the one supportive paper in an evidence set full of contradicting ones. To catch this, Syno runs a separate validation pass after the primary judgments are in. The pass targets every claim whose final verdict is grounded but cites only one paper (the structural shape that lets a cherry-pick survive the first round) and fetches a small set of top uncited candidates from the same retrieval round, filtered to a minimum relevance threshold relative to the cited paper. If a supermajority of those candidates contradict the claim, the claim is demoted and a disconfirming-uncited-evidence flag is attached to the verdict.

A deliberate asymmetry: **silence is not counted as disagreement.** An uncited candidate that returns "not grounded" does not contribute to the demotion count; only an active contradiction does. Keyword-driven retrieval routinely surfaces papers that match the surface form of the query but report on a different population or outcome, and counting their silence as dissent would generate spurious demotions on exactly the topics where retrieval recall is weakest. The pass is also time-bounded, with a budget cap per review.

At retrieval time, Syno reserves dedicated places for papers methodologically adjacent to the query but topically off the obvious answer, the dissenting evidence that ranking-by-relevance would otherwise filter out. These candidates compete on relevance alongside ordinary search hits, so the reservation does not lower the bar for them; it only ensures they reach the candidate set in the first place. This mirrors the discipline a careful human reviewer applies when asking *what does the rest of the literature say?*, and it is rarely implemented in current AI literature-review tools.

2.5 Multi-tier, cross-family escalation

The validation chain is organised in tiers. A fast first-pass tier handles the bulk of judgments; when its confidence is low, when its output disagrees with the drafter's intent in specific ways, or when prior rounds have flagged a claim as contested, the same input is escalated to a deeper-reasoning model with a thinking budget. Escalation is a structured re-judgment by a model with different inductive biases, not a simple retry.

Two non-obvious properties matter for trust:

1. **Cross-family fallback.** Drafter and validator fallback chains explicitly cross model families and vendors. A single-provider outage, a model regression, or a vendor policy change cannot silently degrade the output: the chain falls back to a different family rather than to a degraded same-family tier. This is a structural property, not a per-incident workaround. The 2025 Cochrane Joint Position [8] explicitly calls out provenance and reliability across AI components as evidence-synthesis prerequisites; cross-family fallback is the corresponding architectural commitment.
2. **Escalation provenance.** Each verdict records whether escalation fired and whether the deeper tier overrode the first-pass judgment; the audit trail captures which tier signed off and how many escalations a claim has accumulated. A first-pass judgment and a deep-reasoning judgment do not carry the same weight, and the record preserves the difference. When the escalation budget is exhausted before a verdict resolves, the claim receives a distinct *escalation-dropped* verdict (§4.1) rather than being silently demoted into the same category as scope-rejected claims. A second escalation trigger fires when the same claim re-emerges across revision rounds in modified wording while pointing at the same evidence base.

2.6 Auditable verdicts with policy versioning

Every Syno verdict carries: cited paper IDs; the passage anchor (paper ID plus the verbatim segment the validator scored against); per-axis judgments; the audit record of escalation activity; and a **policy version** stamp identifying the scoring regime in force.

When the scoring regime changes (new weights, new axes, new thresholds), the version bumps. Verdicts stamped with an older version are not reused; they are re-judged. This mirrors the discipline PRISMA requires of a published search: the date and conditions of evaluation are recorded, not assumed. A literature review whose scoring logic is undated cannot be reproduced.

3. Methodological Lineage

Syno sits inside the established lineage of evidence-synthesis methodology, and we want the lineage visible. The reporting and quality-appraisal canon predates the AI era; the AI-specific frameworks have arrived only recently. We engage with both layers explicitly.

Standard	Function in the field	Syno's mechanism
PRISMA 2020 [4]	Reporting checklist for systematic reviews; mandates reproducibility of the search	Fixed evidence set identified by a recorded cryptographic fingerprint, policy-version stamp, audit-trailed verdict record, PRISMA 2020 four-stage flow emitted on every run.
PRISMA-S [supp.]	Extension for reporting the search strategy itself: search strings, databases, filters, date limits	Reproducibility of the evidence base under fixed inputs is supported by the cryptographic fingerprint of the corpus; PRISMA-S–compliant search-strategy reporting (search strings, databases, filters, date limits) is on the roadmap (§8).
Cochrane Handbook [5]	Protocol-first scope, dual-reviewer screening, quality-weighted synthesis, risk-of-bias appraisal	Quality-weighted aggregation (§2.3); escalation as a dual-extractor–with–adjudicator pattern (§4.4); formal per-study risk-of-bias appraisal is on the roadmap.
GRADE [6]	Certainty-of-evidence rating across five down-rating and three up-rating domains	Per-axis outputs surface inputs GRADE assessors require: population fit (one component of indirectness) and outcome directness. Numerical fidelity is treated as a precondition to GRADE imprecision rating, not a substitute for it.
SANRA [7]	Six-item editor-facing quality scale for narrative reviews; rewards explicit search description and faithful endpoint reporting	Per-axis outcome alignment; recorded abstention operationalises the explicit acknowledgement of evidence gaps SANRA's data-presentation and referencing items reward; passage anchors.
Cochrane–Campbell–JBI–CEE Joint Position 2025 [8]	The first cross-organisation position on AI in evidence synthesis; calls for transparency, reproducibility, and human oversight	Policy-version stamping; passage anchoring; the audit trail; explicit abstention; cross-family fallback as a reliability commitment.
PRISMA-trAIce [9]	A proposed reporting checklist for AI-assisted evidence synthesis (JMIR AI, 2025)	Per-claim verdict records, escalation provenance, and corpus-manifest disclosure correspond to the checklist's transparency themes; a row-by-item mapping is in preparation.

The 2019 work of Marshall and Wallace on systematic-review automation [12] is the canonical precursor to the AI-era frameworks. We treat it as the foundation the post-2024 reporting frameworks build on.

We name these standards explicitly because the first generation of AI literature-review tools has not. A literature-review agent that does not engage with PRISMA, Cochrane, GRADE, SANRA, the 2025 Joint Position, and PRISMA-trAIce is, in our view, closer to a summariser of search results than to a literature-review instrument.

4. Trust and Verification Posture

A literature-review system is a piece of clinical and scientific infrastructure. The trust posture matters at least as much as the prose quality. Syno's posture has five pillars.

4.1 Per-claim verdicts, not document-level confidence

Most AI tools, when they produce confidence at all, do so at the document level: a single "trust me" score on the whole output. That score is uninformative, because a 7,000-word review contains hundreds of claims of widely varying defensibility, and an overall number tells the reader nothing about which paragraph to scrutinise.

Syno emits a multi-level verdict record for **every claim**, distinguishing grounded, weakly grounded, not grounded, contradicted, literature-disagrees, and abstain states. Each carries the per-axis judgments that produced it.

Verdict	One-line meaning
Grounded	Cited paper(s) directly support the claim along all relevant axes; cross-paper check passes.
Weakly grounded	Support exists but is qualified (e.g. broader population, surrogate outcome, or single small cohort).
Not grounded	No retrieved paper substantively supports the claim.
Contradicted	At least one cited paper actively conflicts with the claim along a substantive axis.
Literature disagrees	Within the retrieved evidence set, multiple credible papers point in opposing directions; dissent is preserved.
Abstain (scope-rejected)	The claim falls outside the locked scope or is trivially unevaluable; not pursued.
Escalation-dropped	The escalation budget was spent before a verdict resolved; kept distinct from scope rejection in the audit trail.

A clinical-research lead reading a Syno output can scan the verdict record, identify the weakly grounded and literature-disagrees claims, and direct scrutiny there, rather than re-reading the entire document with equal suspicion. The distinct escalation-dropped verdict preserves the difference between "we did not pursue this" and "we tried and could not resolve it," which matters for downstream auditing.

Two further signals travel with the verdict where they apply. For literature-disagrees claims, a temporal direction is recorded (*newer evidence supports*, *newer evidence contradicts*, or *same-era disagreement*), so the reader can see whether the disagreement is a live debate or a generational shift. For any claim with dissenting evidence in the retrieved set, the specific dissenting papers are surfaced rather than collapsed into a single counter. *"Three papers support, one dissents"* is more useful than *"weight-of-evidence supports"* when the dissent comes from the largest or most recent trial.

4.2 Passage anchoring

Every supporting verdict carries a passage anchor: a paper ID and the verbatim segment text the validator scored as entailing the claim. A reader can open any grounded claim and see the exact sentence(s) that justify it: the specific passage, rather than a pointer to "the abstract" or "the methods section." This is the verification artefact competitor users have explicitly asked for ("verbatim quote snippets" recurs in the SciSpace and Elicit feedback corpora).

Passage anchoring is treated as a structural commitment, not a post-hoc decoration. The validator is required to commit to the specific row, subgroup, and verbatim text it is grading against, together with the topic the claim asserts and the numerical content the claim contains, before it produces a judgment label. Self-inconsistency between what the validator committed to and what it concluded is recorded and treated as a quality failure rather than ignored.

4.3 Version-stamped verdicts and reproducibility

Every verdict carries the policy version under which it was computed. When the scoring regime changes, the version bumps and prior verdicts are re-judged on the next run. Combined with the recorded cryptographic fingerprint of the evidence set, this makes Syno's outputs reproducible in the PRISMA-S sense at the level of the evidence base: a third party can re-execute the review and verify that the same evidence, judged under the same policy, produces the same conclusions.

4.4 Escalation as a first-class signal

When the first-pass tier and a deeper escalator tier disagree, Syno neither silently averages nor discards the cheaper opinion. The disagreement is recorded; the deeper tier's verdict is the one carried forward, tagged as having required escalation. Persistent escalation pressure across draft-and-revise rounds signals that the claim is contested in the literature, and that contestation is surfaced in the revision hint. This is analogous in spirit to the dual-extractor-with-adjudicator pattern Cochrane prescribes for data extraction (Handbook Ch.5) [5], where two independent judgments reduce single-source error whether the judgments are human or machine; cross-family fallback (§2.5) extends the same principle to the model-vendor dimension.

4.5 Auditable abstention

When evidence is insufficient, Syno abstains explicitly and surfaces the gap rather than silently omitting it. Competitor users have repeatedly asked for "information not available" honesty; SANRA's data-presentation and referencing items reward explicit acknowledgement of evidence gaps [7], and recorded abstention with a stated reason operationalises that requirement. The distinct *escalation-dropped* verdict (§4.1) preserves the difference between "this was outside the locked scope" and "this was attempted and could not be resolved within the available reasoning budget."

The same commitment applies when prerequisite stages fail before validation can occur. In that case the system declines to emit a review rather than presenting an unvalidated draft as one. A validation layer that is structural rather than cosmetic must behave this way: in the absence of a validated claim record, no review is emitted.

5. System Architecture at a Glance

Syno's review pipeline runs as a sequence of well-defined phases (Figure 1). Each phase has a single responsibility and emits an artefact reproducible from its inputs alone.

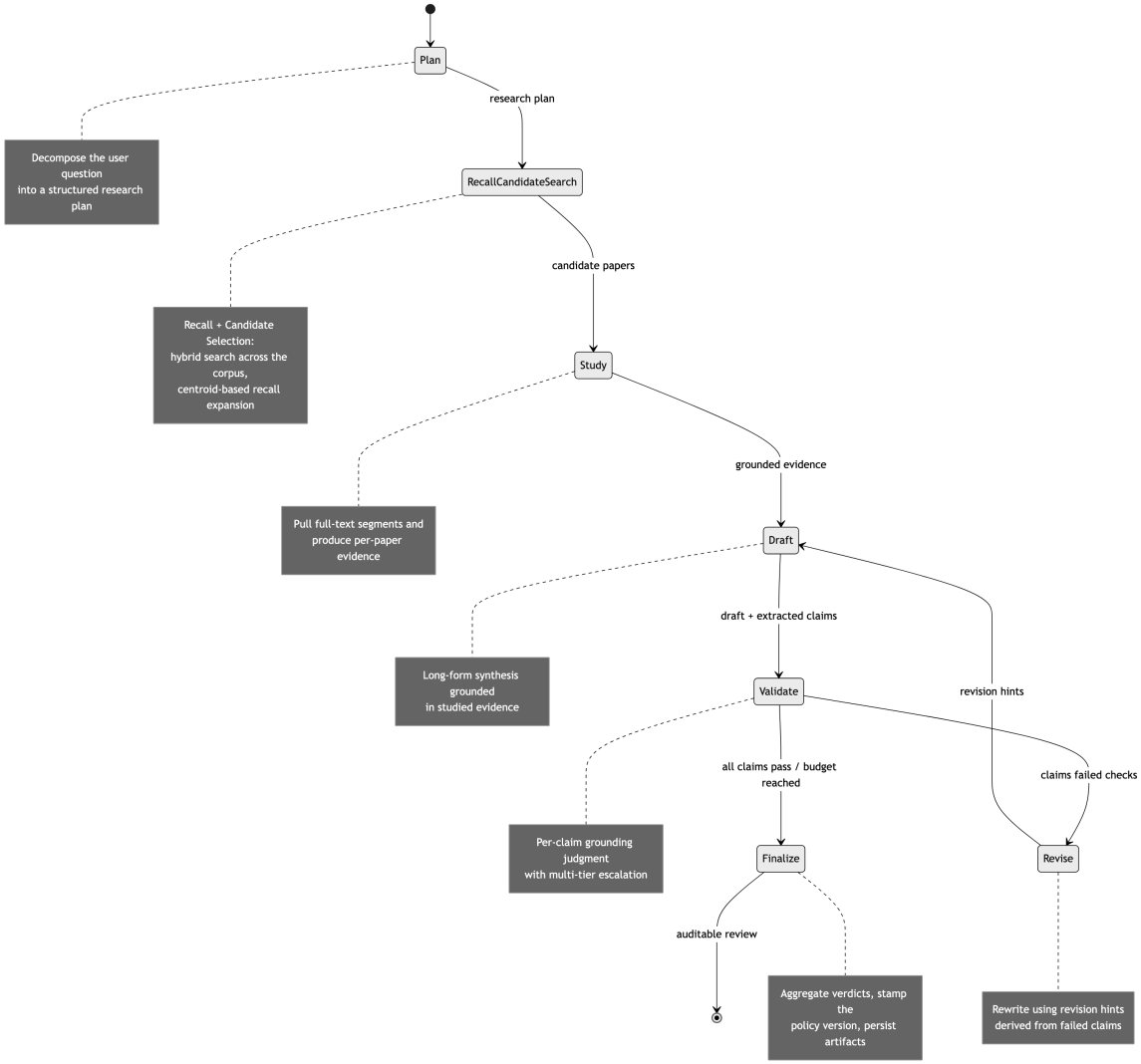

Figure 1. The Syno review pipeline. Each phase produces a typed artefact for the next; failed claims return from validation to revision, and the cycle terminates when the per-claim grounded rate clears its target or the budget is reached.

The pipeline begins with **planning**: the user question is decomposed into a research plan, with PICO-style framing where applicable [4], inclusion criteria, and the review posture (systematic / scoping / narrative) recorded alongside the active policy version.

The **retrieval** phase queries a curated biomedical corpus indexed over the identifier space researchers already work in (PMID, PMCID, DOI), with arXiv added to capture preprints in fast-moving therapeutic areas. Hybrid retrieval pairs a dense biomedical-embedding channel with a precision-oriented lexical channel and re-scores the result; a cross-encoder reranker is available for high-recall and benchmark modes. Every run records a cryptographic fingerprint of the corpus it queried, so a third party can confirm two runs were executed against the same evidence base. A dedicated reserve of slots protects dissenting-axis evidence (§2.4) from being filtered out by a relevance ranking dominated by the answer's surface form.

In the **study** phase, each retained paper is processed into a structured metadata record (study design, population, sample size when stated) plus a faithful multi-paragraph summary. These artefacts feed both drafting and validation.

In the **drafting** phase, a long-context reasoning model composes the prose; every substantive claim is tagged with the supporting paper identifiers from the retained evidence set. The writer cannot reference a paper outside that set.

The draft then enters the **per-claim validation** phase. Free prose is first decomposed into an atomic per-claim record (one entry per assertion, with the papers the writer cited) because long stacked sentences that conceal partial-truth failures (§1.3) cannot be judged as a whole. Each claim is then graded along the per-axis dimensions of §2.2, weighted-aggregated per §2.3, run through the cross-paper dissent check of §2.4, and escalated per §2.5 when contested. Every model call carries

a wall-clock budget; a review-level limit caps total wall time so a stuck escalation cannot consume the whole compute budget.

In the **revision** phase, the writer is shown structured verdicts, reframing hints, and, for claims demoted to weakly grounded or contradicted, concrete replacement candidates from within the same evidence set, each accompanied by its full summary. Population mismatch becomes an invitation to narrow the prose; a contradicting body of evidence surfaces as alternative citations the writer can swap in, not just a rejection signal. Already-grounded claims are held stable across revision rounds so that work the prior round verified is not silently re-judged. When evidence is insufficient, the system abstains explicitly.

Finally, the **audit-output** phase emits the complete review: prose, per-claim verdict record, passage anchors, policy version, corpus fingerprint, a PRISMA 2020 four-stage flow (identified → screened → assessed → included), and a per-paper bibliography table carrying study-design, citation-count, and relevance-score provenance, including papers referenced only in the prose without a validator verdict, labelled as such. When the time limit fires before the last revision round completes, the system recovers to the strongest prior round rather than emitting a regressed final-round result; the recovery is recorded so a reader can see what happened. When no round produced verdicts at all, the review is withheld rather than emitted on default assumptions.

A reader can regenerate the output from these inputs alone, or audit any single sentence against the passage that justified it.

6. Design Rationale and Internal Testing Posture

This section describes how Syno's own validation chain is calibrated; the external evaluation lives in §7.

What we test for. Internal evaluation is organised around the failure modes in §1: per-claim citation faithfulness (does the cited paper support the surrounding sentence, axis by axis), recall on dissenting evidence (does the cross-paper check catch a single-supporter cherry-pick), and behaviour under known-hard cases (very recent topics, sparse evidence bases, "current consensus" claims when consensus has shifted). The validation chain is calibrated against these properties, not against aggregate fluency.

How we test. We maintain internal gold sets (biomedical questions paired with curated claim-level expected behaviour) built to include cases the system is expected to handle poorly alongside cases it should handle well. Gold-set construction is itself version-stamped; gold sets are not silently rewritten to make the system look better. Calibration optimises for per-claim citation faithfulness as the primary signal, with claim recall and synthesis quality secondary.

Self-graded calibration vs external benchmark. We deliberately separate two epistemic objects. *Internal calibration*, Syno's own validator scoring Syno's own output across rounds, is what tunes the validation chain; it shows the validation-and-revision cycle closing on grounding, and is reported (with the explicit self-graded caveat) in Annex A.6. *External benchmark* (§7), Syno's final output scored by independent judges that did not see Syno during training or design, is the one that may legitimately be compared with other tools. Conflating the two is a common error in this field; we keep them on different pages.

7. Results: External Benchmark (Diabetes)

One run, one topic family, one corroborating data point. This is not the only evidence the architecture stands on, and we treat it as one piece rather than the thesis. The full method, all caveats, the 9 grader-bug corrections we made (including ones that favoured Syno and we fixed against our own interest), and the downloadable raw verdicts live in Annex A. The numbers below are computed by majority of three independent judges per (claim, paper) unit before any rate is taken; the inferential statement is at the question-cluster level.

What we ran. Ten diabetes systematic-review questions (q01–q10), one capture per tool. The comparison set spans the two categories a biomedical researcher actually reaches for: **Syno** against two purpose-built AI literature-review assistants (**Elicit** and **Consensus**) and one general-purpose deep-research agent that is not literature-review-specific, **Gemini Deep Research** (included because general agents are widely used for this task in practice). Each tool's output was put through the *identical* pipeline: source-attribution cues stripped, per-citation existence-verified against Crossref / NCBI E-utilities / OpenAlex / Unpaywall (deterministic, no LLM in the loop), then per-claim faithfulness graded by an external panel of three judges drawn from two model families (Anthropic Claude Sonnet 4.6 + Google Gemini 2.5 Pro + Gemini 2.5 Flash) against the actual fetched text of each cited paper. The primary endpoint, the exclusion rules, and the analysis plan were pre-registered before the run.

Primary endpoint (Figure 2). On per-claim numeric fidelity (whether the cited paper actually contains the number the prose attributes to it), Syno led Elicit by a per-question mean of **+18.8 percentage points** (95% t-CI [+9.4, +28.2] across the K=10 question clusters), and was ahead in **9 of 10 questions**. The CI excludes zero, and the lead is not driven by a single question.

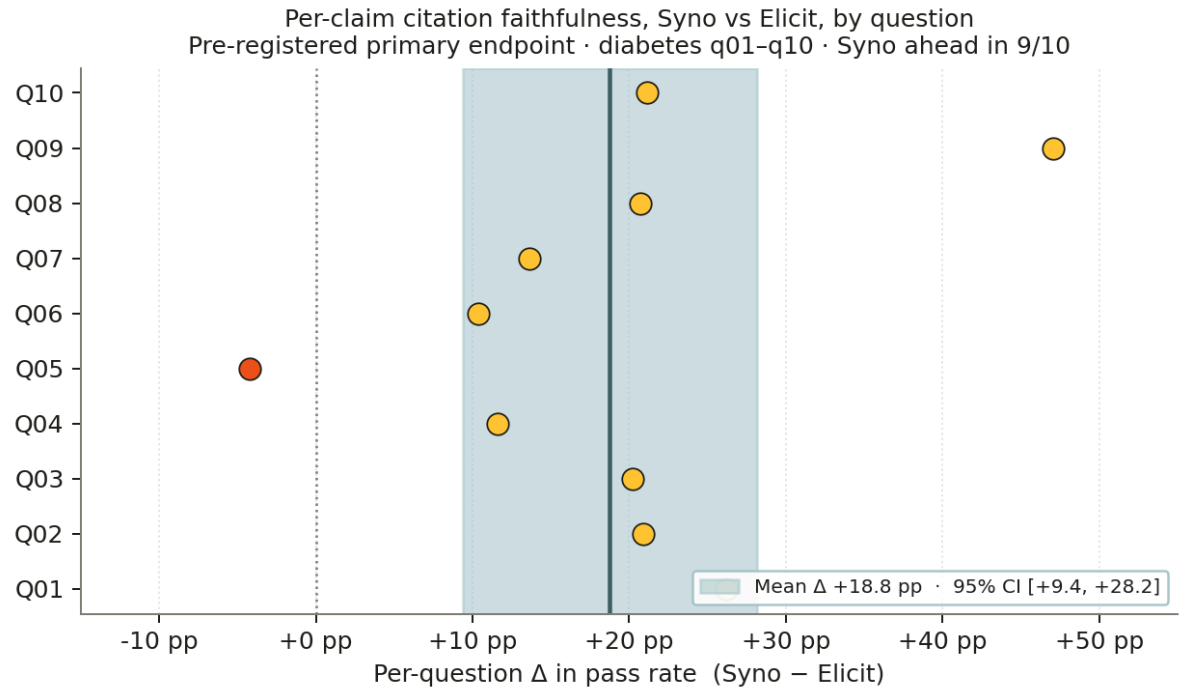

Figure 2. Per-question Syno − Elicit Δ on per-claim numeric fidelity. Each marker is one question; the band is the 95% confidence interval on the mean per-question delta. Syno ahead in 9/10. Pre-registered primary endpoint.

Citation existence (the trust floor). On the deterministic citation-existence layer, Syno (100.0%), Consensus (99.5%), and Elicit (98.4%) cluster at the top with effectively all citations resolving to indexed scientific papers. (These three expose only author–year markers, not source URLs, so for them only *resolvability* is measurable, not source type.) **Gemini Deep Research sits well below the others here**: 69.2% of its citations resolve, with **12.1% pointing to non-scientific sources** (press releases, ClinicalTrials.gov, Scribd, RxFiles, consumer-health blogs) and a further 18.7% well-formed but unverifiable. The non-scientific-source share is observable only for Gemini because it is the one tool that exposes its source URLs. As a general-purpose deep-research agent rather than a purpose-built literature tool, it is not citing from the same source mix as the others in the first place.

Robustness to the judge. Syno's relative lead is not an artefact of any one judge family. Computed on the Claude judge alone the Syno − Elicit gap is +13 percentage points; on the two Gemini judges alone it is +16; on the majority-of-three collapse it is +17. (The two Gemini-family judges are uniformly more lenient than Claude on every tool by 17–25 pts, including, notably, on the Gemini tool itself, which is the opposite of self-preference. We read this as a *calibration spread*, not a conflict of interest; the per-family table is in Annex A.4.)

Against the human baseline (loose context). The careful-human per-claim quotation-accuracy band is ~75–83% [13][14][15]. Syno's external faithfulness rate *brackets* that band: the strict Claude-only read (69% pooled) sits just below it, the lenient majority panel (92% pooled) at or above it. These are different measurements: an LLM panel judging an AI tool's claim faithfulness is not the same thing as human authors' quotation accuracy, so we use the band only as a rough scale check, not as a claim that Syno matches human reviewers.

Honest scope and trade-offs. The result is corroborating, not conclusive. (i) Diabetes only: cross-domain generalisation is future work. (ii) Single capture per tool, so run-to-run stability is not measured here. (iii) An external *LLM* panel rather than a human expert panel, with a complementary ≥3 human biomedical rater study acknowledged as future work. (iv) **Full-text vs abstract.** Syno was graded on 100% full-text; Elicit and Gemini mostly on abstracts (each tool's actual evidence base). This is not a criticism of their design choice: full-text redistribution licences from the major biomedical publishers are economically unrealistic at consumer pricing, and abstracts-plus-open-access is a rational response. Syno takes the opposite trade-off (narrower coverage for full-text grounding) by design. We grade each tool on the evidence base it actually serves in production, a deliberate *deployed-as-used* comparison rather than handicapping Syno to an abstract-only run; full-text grounding is the point of the product and, in our view, the higher-quality evidence, so the result reflects evidence base and method together, stated plainly. (v) Syno's multi-round validation loop runs ~2× slower than single-pass tools, an accepted trade-off; the relevant comparison is to a careful human review, against which Syno is dramatically faster. (vi) Syno's corpus is currently *smaller* than the competitors': diabetes-scoped, full open-access only, by design. The diabetes scope is an operational choice, not an architectural one, and the domain-agnostic pipeline is expected to carry to other disease families. Coverage breadth is a real limitation today; broadening it without dropping the full-text standard is ongoing work. The full 8-item caveat list is in Annex A.5.

Verify it yourself. The downloadable data package (syntactiq.ai/literature/whitepaper/reference-data.zip) contains every question, every tool's raw graded output, the per-judge verdicts (CSV), the judge prompts, and the cross-validated review record. Anyone is welcome to re-grade with their own judges or method.

8. Limitations and Roadmap

Honest accounting of what Syno does **not** yet do:

Reporting artefacts. Each review emits a PRISMA 2020 four-stage flow and a per-paper bibliography table (§5). Not yet generated: GRADE Summary-of-Findings tables, formal risk-of-bias traffic-light plots, and evidence-gap maps. Not blocked on architecture; blocked on prioritisation.

Search-strategy provenance. Queries are not yet constructed from MeSH-controlled vocabulary blocks joined with Boolean operators, and we do not yet emit a PRISMA-S–compliant search-strategy artefact. This requires deeper integration with PubMed E-utilities and Embase Emtree licensing. It is among the most consequential gaps against generalist tools and we are scoping it.

Determinism. Syno's evidence layer is reproducible by design: paper selection and retrieval are deterministic on our literature server and are pinned by the recorded cryptographic fingerprint of the evidence set, so the same question over the same corpus returns the same papers. The only residual variation is the token-level sampling of the underlying reasoning models, a property shared by every LLM-based tool in this comparison rather than a Syno-specific limitation. For workflows that need bit-identical reruns, a stricter determinism mode (temperature pinning, fixed seeds, recorded model versions) is in development.

External evaluation (scope). This version of our benchmark (§7, Annex A) is diabetes-only, single capture per tool, graded by an external LLM panel. It covers four systems: Syno, two purpose-built literature-review tools (Elicit, Consensus), and one general-purpose deep-research agent (Gemini Deep Research). Cross-domain generalisation (cardiology beyond diabetic complications, oncology, neurology), a multi-run stability sub-analysis, and a complementary ≥3 human biomedical rater panel are future work, as is a broader tool field (for example Perplexity Deep Research, OpenAI Deep Research, OpenEvidence, Undermind). OpenEvidence in particular stopped serving the European region in which we operate, so it could not be included; we will revisit once access returns.

Full-text access. Retrieval surfaces PubMed/PMC plus openly available content; paywalled full-text from major publishers is not yet integrated. Unpaywall and institutional-proxy integration are on the roadmap.

Living-review updates. Syno produces one-shot reviews. The infrastructure to monitor the literature on a cadence and emit a *what changed* delta is roadmap work; the policy-versioning architecture is designed to support it.

Risk-of-bias appraisal. No formal RoB 2 or ROBINS-I per study yet. The aggregator uses study-design class as a quality proxy, defensible as a first phase but not a substitute for domain-by-domain RoB 2.

Quantitative pooling. No network meta-analysis, forest plots, or aggregate pooling. Requires extraction quality (effect sizes, CIs, denominators) that we treat as a future milestone.

Non-English literature. Retrieval is English-biased upstream; translation-augmented retrieval introduces a new hallucination surface, deferred until retrieval surfaces meaningful non-English yield.

Domain breadth. Internal calibration is heavily weighted toward endocrinology and metabolic-disease questions where our gold sets are most developed. Generalisation to oncology, cardiology, and neurology is a near-term priority.

Corpus reach. Syno's indexed corpus is currently smaller than the corpora the generalist tools draw from: it is *scoped* (diabetes-focused, in this evaluation) and restricted to *full-text open-access* papers, a deliberate choice in favour of grounding faithfulness over coverage breadth. **Nothing in the architecture is diabetes-specific.** The scope is an operational choice that concentrates effort where Syntactiq currently works most (diabetes), not an architectural constraint: the same retrieval, grounding, and validation methodology applies unchanged to any disease family. Cross-domain results are not yet in, but the pipeline is domain-agnostic by construction, and we are confident it will carry over as we open it up. Widening the corpus while keeping the full-text standard (Unpaywall, institutional-proxy, additional disease families), under the identical methodology, is the active work.

Latency / cost. Syno's per-claim validation-and-revision loop runs roughly twice as long as single-pass deep-research tools. We treat this as an accepted, deliberate trade-off: the relevant comparison for a literature review is to a careful human, against which Syno is dramatically faster. We do not present this as a measured cross-tool benchmark figure; it is an architectural property of running per-claim validation rather than asserting a single document.

We list these gaps explicitly because a tool that hides what it cannot yet do has not earned a careful reviewer's trust.

9. Conclusion

The §1 evidence is concrete. On biomedical question-answering, roughly half of the citations produced by leading generalist deep-research products are fabricated; the one tool that gets existence largely right still produces cited statements in which over half contain at least one subtle inaccuracy or misrepresentation of the source. Structured-extraction tools, even when the cited papers are real, fail mostly through interpretation error rather than citation-existence error. That is the floor the current field leaves for clinical and translational research to absorb.

This is what those failure rates look like under Syno's posture. Retrieval-first grounding structurally prevents out-of-corpus citation fabrication, so the 47–50% fabricated-reference number has no structural place in the output. Per-axis judgment makes misinterpretation legible at the claim level, so the over-50% rate of subtle inaccuracy in deep-research cited statements does not survive un-flagged through validation. The cross-paper dissent check, with deliberate preservation of methodologically adjacent evidence, addresses the dominant Elicit-style failure: a real paper, summarised in a way the rest of the evidence set would dispute. Quality-weighted aggregation displaces vote-counting. Multi-tier, cross-family escalation surfaces contestation rather than averaging over it. Policy-versioned, passage-anchored verdicts make the review something a research librarian can scrutinise, and disagree with, at the level of individual claims. The diabetes-scoped

external benchmark in §7 supplies the first corroborating data point: on this slice, Syno shows a per-claim faithfulness lead over the strongest peer that holds across every judge subset and is fully reproducible from the published data package.

What distinguishes Syno. The property that distinguishes Syno is **auditable per-claim trust**: passage-anchored verdicts, version-stamped policies, escalation provenance, distinct abstain categories, cross-family fallback. That trust is bought deliberately: Syno runs an iterative draft-validate-revise loop and accepts more wall-clock time and compute than single-pass tools, on the view that for this user a slower, costlier, more accurate review is the right trade. Retrieval-recall and workflow-integration plays are valuable and other teams are advancing those frontiers. The question Syno is built to answer is the one a clinical-research lead, an IND filer, or a guideline author needs answered at the point of decision: *can I prove this specific sentence is what the literature actually says?* The architectural choices behind that answer are commitments built in from the start, not wrapped around an existing system afterwards. For clinical and translational researchers whose downstream decisions are not abstract, that is the kernel that matters.

References

[1] Wong MYH, Ong AY, Merle DA, Keane PA. Deep Research Agents: Major Breakthrough or Incremental Progress for Medical AI? *Journal of Medical Internet Research* 2026;28:e88195. DOI: 10.2196/88195. https://www.jmir.org/2026/1/e88195

[2] Jaźwińska K, Chandrasekar A. *AI Search Has A Citation Problem.* Tow Center for Digital Journalism, Columbia Journalism Review, 2025. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php

[3] Lagisz M, Mizuno A, Morrison K, Pollo P, Ricolfi L, Yang Y, Nakagawa S. Using Elicit AI research assistant for data extraction in systematic reviews: a feasibility study across environmental and life sciences. *EcoEvoRxiv* preprint, 2025. DOI: 10.32942/X2F346. See also: Hilkenmeier F, Pelzer M, Stierle C, Fink-Lamotte J. Evaluating the AI Tool "Elicit" as a Semi-Automated Second Reviewer for Data Extraction in Systematic Reviews: A Proof-of-Concept. *Social Science Computer Review* 2025 (OnlineFirst). DOI: 10.1177/08944393251404052.

[4] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. DOI: 10.1136/bmj.n71. https://pmc.ncbi.nlm.nih.gov/articles/PMC8005924/

[5] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions*, version 6.5. Cochrane, 2024. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current

[6] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924–926. DOI: 10.1136/bmj.39489.470347.AD.

[7] Baethge C, Goldbeck-Wood S, Mertens S. SANRA — a scale for the quality assessment of narrative review articles. *Research Integrity and Peer Review* 2019;4:5. DOI: 10.1186/s41073-019-0064-8. https://pmc.ncbi.nlm.nih.gov/articles/PMC6434870/

[8] Flemyng E, Noel-Storr A, Macura B, et al. *Position Statement on AI Use in Evidence Synthesis Across Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence 2025.* Campbell Systematic Reviews 2025. DOI: 10.1002/cl2.70074.

[9] Holst D, Moenck K, et al. Transparent Reporting of AI in Systematic Literature Reviews: Development of the PRISMA-trAIce Checklist. *JMIR AI* 2025;e80247. DOI: 10.2196/80247.

[10] Lau O, Golder S. Comparison of Elicit AI and Traditional Literature Searching in Evidence Syntheses Using Four Case Studies. *Cochrane Evidence Synthesis and Methods* 2025. DOI: 10.1002/cesm.70050.

[11] Bianchi F, et al. Data Extractions Using a Large Language Model (Elicit) and Human Reviewers in RCTs. *Cochrane Evidence Synthesis and Methods* 2025. DOI: 10.1002/cesm.70033.

[12] Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews* 2019;8:163. https://pmc.ncbi.nlm.nih.gov/articles/PMC6621996/

[13] Baethge C, Jergas H. Systematic review and meta-analysis of quotation inaccuracy in medicine. *Research Integrity and Peer Review* 2025. DOI: 10.1186/s41073-025-00173-z. https://pmc.ncbi.nlm.nih.gov/articles/PMC12285159/

[14] Jergas H, Baethge C. Quotation accuracy in medical journal articles — a systematic review and meta-analysis. *PeerJ* 2015;3:e1364. https://pmc.ncbi.nlm.nih.gov/articles/PMC4627914/

[15] Mogull SA. Accuracy of cited "facts" in medical research articles: A review of study methodology and recalculation of quotation error rate. *PLOS ONE* 2017;12(9):e0184727. https://pmc.ncbi.nlm.nih.gov/articles/PMC5599002/

Supplementary methodological references (cited in passing):

- Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898.
- Sterne JAC, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919.
- Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews* 2021;10:39.
- Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews. *BMJ* 2017;358:j4008.

Annex A. Benchmark: method, results, caveats, corrections

This annex is the bridge between the §7 headline and the downloadable data package. It restates the method in compact form, gives the full results tables, lists every caveat, and lays out the bugs we found in our own grader. Detail beyond this (every prompt, every captured tool output, every per-judge verdict) lives in the data package zip.

A.1 Scope and pipeline

Ten pre-registered diabetes systematic-review questions (q01–q10); one capture per tool; four tools (Syno, Elicit, Consensus, Gemini Deep Research). Each tool's output passed through one shared pipeline:

Layer	What it does	LLM in the loop?
L0 — Capture & blind	Each tool's output is normalised to one schema; source-attribution and tool-identity cues scrubbed.	No
L1 — Citation resolution	Every cited reference verified against Crossref / NCBI E-utils / OpenAlex / Unpaywall. Outcomes: *resolved* · *non-scientific source* · *unverified* · *parse-failure (excluded).*	**No — deterministic, $0-LLM, fully reproducible.**
L2 — Source-text retrieval	The cited paper's actual text fetched (full text where available, abstract otherwise). Same cascade for every tool's paper IDs, regardless of which tool cited them.	No
L3 — Per-claim faithfulness panel	Three judges from two model families (Anthropic Claude Sonnet 4.6 + Google Gemini 2.5 Pro + 2.5 Flash) grade each (claim, paper) unit along five axes: numeric, overall support, population, outcome, recency.	Yes (disclosed; with safeguards)
Aggregation	Per unit, the three judge verdicts are collapsed by **majority of three** before any rate is computed. Inferential statement is clustered by question.	No

How a rate is computed (so the tables reproduce from the published verdicts). For each (claim, cited-paper) unit the three judges' verdict on an axis is collapsed by **majority of three**; a unit a judge marks *not-assessable*, or whose source text was unavailable, is **excluded from that axis's denominator**, identically for every tool. A unit **passes** an axis when its collapsed verdict is *supported* (*weakly-supported* and below do not pass). Each cell is passes ÷ assessable units with a Wilson 95% interval; the pre-registered primary endpoint's interval is instead clustered by question (ten clusters). Applying this rule to 04_data/verdicts.csv regenerates the Annex tables.

The full method, judge prompts, and pre-registration are in benchmark-evaluation-protocol.md, benchmark-prereg.md, and benchmark-validation-overview.md (all included in the data package).

A.2 Citation resolution: full table

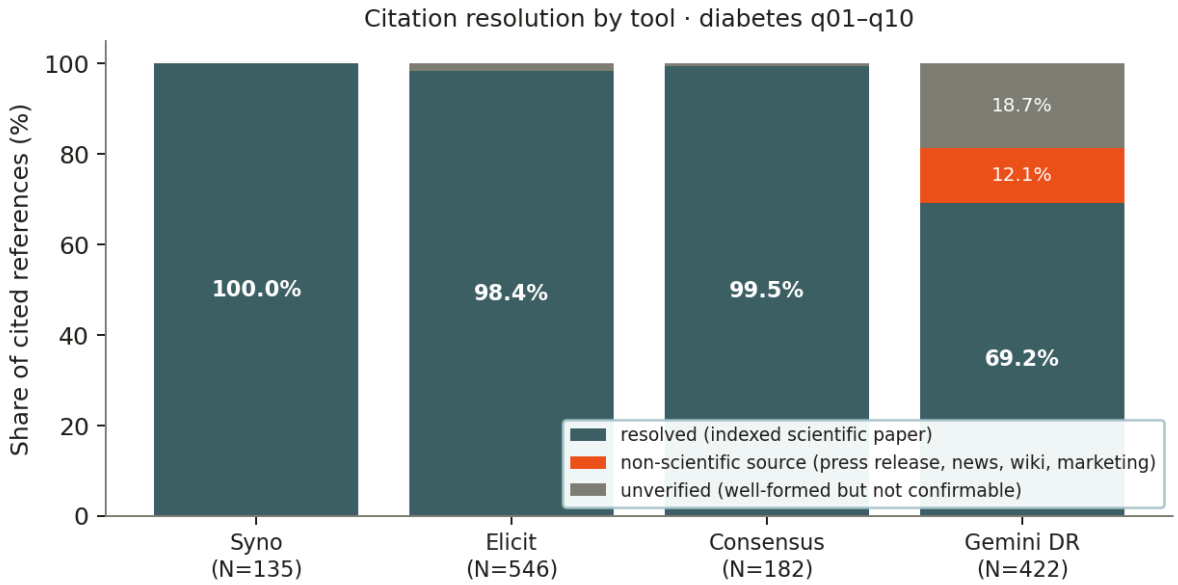

Figure 3. Citation resolution by tool, q01–q10. *Resolved* = the cited reference exists in an authoritative scientific index (Crossref / NCBI / OpenAlex). *Non-scientific source* = the cited URL is a press release, news article, wiki, marketing page, or other non-peer-reviewed item. *Unverified* = the reference is well-formed but cannot be confirmed against any authority.

Tool	N references	Resolved	Non-scientific	Unverified
Syno	136	**100.0%**	0.0%	0.0%
Consensus	184	99.5%	0.0%	0.5%
Elicit	547	98.4%	0.0%	1.6%
Gemini Deep Research	422	69.2%	**12.1%**	18.7%

Rates are over the gradeable denominator (N − parse-failure); parse-failures (1–2 per tool) are our parser's fault and never counted as a tool demerit.

A.3 Per-claim faithfulness: full per-axis table

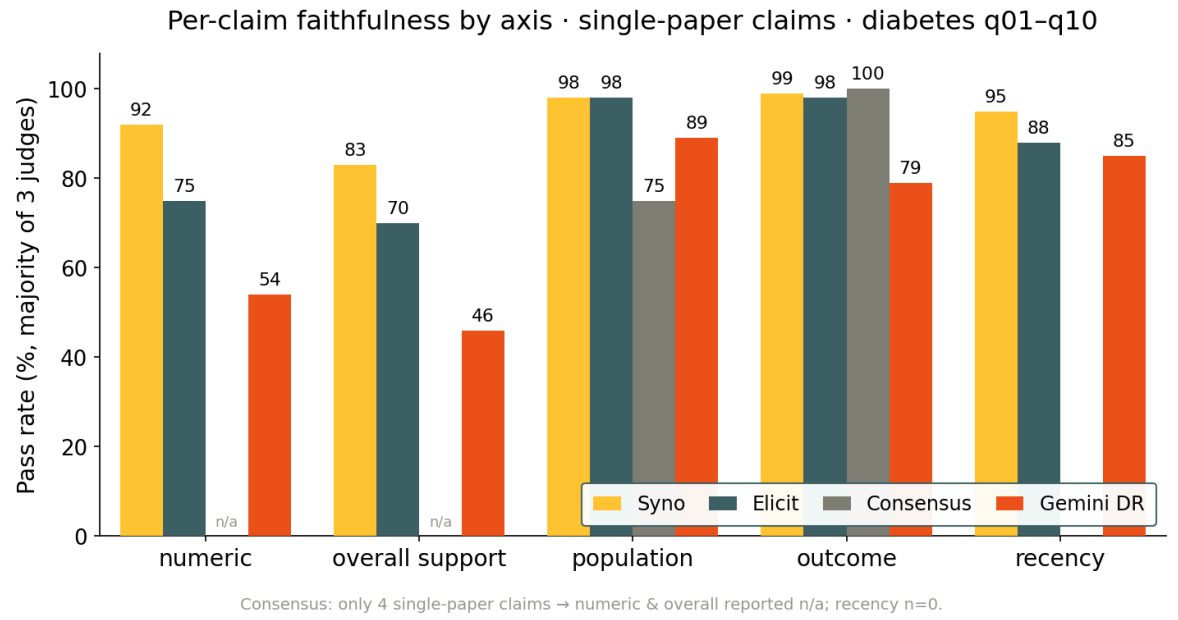

Figure 4. Per-claim faithfulness by axis and tool, q01–q10. Pass rates are computed as the majority of three judges per (claim, paper) unit; only single-paper claims are included so a number is cleanly attributable to one paper. *Population* and *outcome* are saturated near 100% across all tools; differentiation lives on *numeric* and *overall support*.

Axis	Syno	Elicit	Consensus	Gemini DR
Numeric fidelity *(primary)*	**92%** (n=280)	**75%** (n=170)	n/a (n=4)	54% (n=104)
Overall support	**83%** (n=295)	70% (n=187)	n/a (n=4)	46% (n=134)
Population	98%	98%	75%	89%
Outcome	99%	98%	100%	79%
Recency	95%	88%	—	85%

Synthesis mode (multi-paper "collective support" claims, where Consensus's synthesis-style writing actually has numeric N): Syno 83% (n=63), Elicit 44% (n=101), Consensus 38% (n=24); Gemini had no multi-paper numeric units.

A.4 Judge-family robustness (the conflict-of-interest check)

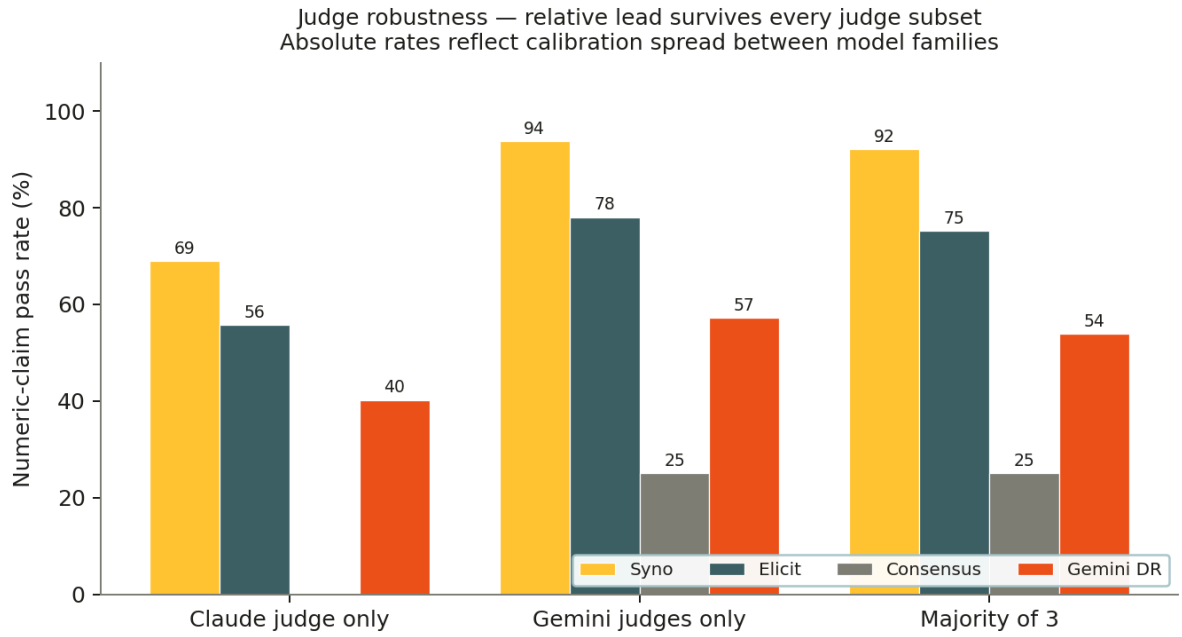

Figure 5. Per-claim numeric fidelity by judge subset. Syno's relative lead is preserved across every subset; absolute rates reflect the calibration spread between model families, not a self-preference effect.

The panel is 2/3 Gemini-family. We measured whether the Gemini judges favour the Gemini *tool*. They do not: the Gemini judges are uniformly more lenient than the Claude judge by 17–25 percentage points across *every* tool, and the gap is *smallest* on the Gemini tool itself (the opposite of self-preference). We read this as **calibration spread**, not conflict of interest, and report all three subsets (Claude-only, Gemini-only, majority-of-three) rather than burying it.

Tool	Claude only	Gemini judges only	Majority of 3
Syno	69%	94%	**92%**
Elicit	56%	78%	**75%**
Gemini DR	40%	57%	**54%**

Inter-rater agreement (Fleiss κ, three judges, single-paper claims): numeric 0.40, overall 0.46, population 0.33, outcome 0.53, recency 0.58. Fair-to-moderate, consistent with the calibration spread.

A.5 Caveats: the full list

Every one of these leans *against* over-claiming for Syno.

1. **Judge-calibration spread.** The pooled 92% absolute number is inflated by the more lenient 2/3 Gemini majority. The defensible claim is the *relative* lead (Syno over Elicit), which holds for every judge subset, not the rosy absolute rate.
2. **Full-text vs abstract, and why each tool's choice is reasonable.** Syno was graded on 100% full-text papers; Elicit and Gemini on ~90–95% abstracts (of graded pairs). Syno's lead is *partly* corpus composition. This is **not a criticism of the other tools' design choice**: commercial AI products at consumer or low-per-seat pricing essentially cannot license redistribution rights for full text from the big biomedical publishers; those licences are negotiated per-

publisher, often six- to seven-figures per year, and they ruin the economics of any service priced for general use. Operating on abstracts and open-access full text is the rational response. Syno takes the opposite trade-off (narrower coverage in exchange for full-text grounding) because per-claim faithfulness is what we are optimising for, not breadth. Either choice is defensible; both should be disclosed, which is what this caveat does. We do not plan a matched abstract-only run: full-text grounding is the product's deliberate choice and, in our view, the higher-quality evidence, so we report each tool as it actually operates rather than handicapping Syno to its competitors' evidence base. The result reflects evidence base and method together, stated rather than normalised away.

3. **Inter-rater κ is fair-to-moderate** (numeric 0.40), not high. Some of the noise is real judgment uncertainty on ambiguous claims.
4. **Consensus single-paper n=4 → reported as N/A** for the per-axis single-paper numbers; its real signal is synthesis-mode (38%).
5. **Gemini's 54% is a conservative floor.** Gemini Deep Research renders many statistics as inline-math images. Its Google-Doc .txt export does preserve those numbers, and ~76% of its numeric claims align to it; the remaining ~24% are sentences our matcher cannot align, so they grade against the PDF (where the statistic is an unreadable image). Gemini's true numeric-fidelity rate is therefore likely *higher* than 54%, and we report the conservative figure rather than inflate it.
6. **Latency / cost.** Syno's multi-round validation loop runs ~2× slower than single-pass tools, an accepted, deliberate trade-off (the relevant comparison is to a careful human, against which Syno is far faster). Stated as an architectural property, not measured here as a cross-tool benchmark figure.
7. **Human baseline (loose context, different construct).** Medical-literature quotation-error rate is ~17–25% (≈75–83% accurate) [13][14][15]. Syno's majority-panel rate (92%) sits at/above that band and its strict Claude-only rate (69%) just below it. Human quotation accuracy and an LLM panel's claim-faithfulness judgement are not the same measurement, so the band is a scale check only, not an equivalence.
8. **Corpus reach.** Syno's corpus is smaller than the competitors': scoped (diabetes-only) and full-open-access-only, by design. Nothing in the architecture is diabetes-specific: the scope is an operational choice (where Syntactiq currently works), not an architectural limit, and the domain-agnostic pipeline is expected to generalise to other disease families under the same methodology. A real coverage limitation today; broadening it while keeping the full-text standard is ongoing work.

A.6 Self-graded calibration trajectory

*The number in this subsection is **self-graded**: Syno's own validator scoring Syno's own output. It is reported here because it motivates the architecture; it is **not** comparable to the external numbers in §7 or in tables A.2–A.4, and we keep it on a separate page deliberately.*

Across a recent window of internal review runs (several hundred claims on biomedical questions), the validator's verdict distribution on the writer's **first pass, before any validation feedback,** sat at approximately four-in-five claims grounded, one-in-ten weakly grounded, and roughly one-in-eleven outright unsupported or actively contradicted *against the very papers the writer just cited*. After the validation-and-revision cycle converged, that distribution shifted to approximately nineteen-in-twenty grounded with the unsupported and contradicted categories each well under one in a hundred. The headline is the *first-pass* number: roughly **one in five claims** from a retrieval-grounded, long-context reasoning model fails the grounded bar before per-claim adjudication. That gap is exactly what the validation-and-revision cycle exists to close.

A.7 Corrections we made to our own grader

The benchmark pipeline went through repeated adversarial review rounds (independent agents, identical task, change made only on convergent findings). We found nine bugs in our own grader and fixed every one, including the three that had been favouring Syno, which we corrected against our own interest.

#	Issue	Bias direction	Fix
1	Blinding leak: a citation-style tell let the panel infer the tool.	Favoured **Syno**	Scrubbed tool tells to a neutral marker; re-graded.
2	Pseudo-replication: counted 3 panel judges as 3 independent observations, inflating n ~3×.	Affected all tools (over-precise)	Collapse to one majority verdict per (claim, paper) before any rate or CI.
3	Panel COI risk: 2/3 judges are Gemini-family while Gemini is graded.	Potential self-preference	Per-judge-family breakdown reported (Annex A.4); shown to be calibration, not COI.
4	looks_tabular over-drop: a comma-count rule discarded legitimate competitor prose.	**Hurt Elicit / Consensus**	Strip citation brackets and author-year groups before counting commas; ~7 legit competitor claims restored.
5	Gemini wrong-number injection (3 fusion classes): sentence fusions could graft a number from a neighbouring clause onto the wrong paper.	**Hurt Gemini**	Hardened sentence-splitting, length-ratio guard, bare-statistic-row drop, dedup; verified 0 wrong-number units across 932 graded units.
6	Resolver under-credited Gemini: stacked source-chrome in titles and a split-subscript artefact (HbA 1c) broke lookup.	**Hurt Gemini**	Strip all trailing chrome; collapse the HbA1c artefact; recovered 4 indexed papers (Gemini 67.8% to 69.2%).
7	native placeholder accepted for Syno citations on the gradeability gate.	Potential favour to **Syno**	Held all tools to the same resolved bar; no live impact (all Syno refs already verified).
8	Orphan verdicts: re-extraction left stale graded rows still counted.	Over-count risk	Report restricted to the current claim set.
9	Internal architecture metric first quoted as 99.6%.	Rosier for **Syno**	Corrected to the honest 95.1% fully-grounded figure; never reverted.

Of these nine, three (4, 5, 6) had been penalising competitors and we fixed them in their favour; three (1, 7, 9) had favoured us and we corrected them against our own interest. That balance is the point of the exercise. The full review record (each round, each independent agent's verdict) is in 05_reviews/ of the data package.

A.8 Data package: anyone can re-grade

A self-contained transparency package is published as a downloadable archive:

Download: syntactiq.ai/literature/whitepaper/reference-data.zip

The archive contains:

- **00_questions/**: the ten pre-registered questions, verbatim.
- **01_methodology/**: the pre-registration (benchmark-prereg.md), the evaluation protocol (benchmark-evaluation-protocol.md), and a plain-language validation overview describing the anti-"AI-grading-AI" safeguards.
- **02_findings/**: the full findings log with every figure, every caveat, and the round-by-round bug-fix history.
- **03_tool_captures/**: every tool's raw graded output (the actual artefact each judge saw).
- **04_data/**: CSV exports: per-citation resolution outcomes (references.csv), per-(claim × paper × judge) verdicts with each judge's evidence quote and rationale (verdicts.csv), and capture metadata.
- **05_reviews/**: the corrections record and the cross-validated review trail.
- **06_figures/**: the four PNGs used in §7 and Annex A.

Deliberately excluded for copyright and size: the fetched full text of cited *source papers* (third-party copyright), and an internal API response cache. Without them you cannot re-run grading locally; with verdicts.csv and its evidence quotes you can audit every verdict we report.

A reader who wants to re-grade with a different judge panel, a different aggregation rule, or a different rubric can do so end-to-end from this package alone; start at syntactiq.ai/literature/whitepaper/reference-data.zip.

Syno is a product of Syntactiq.