TripleCheck: Transparent Post-Hoc Verification of Biomedical Claims in AI-Generated Answers

Ana Valeria González and **Sidsel Boldsen** and **Roland Hangelbroek**
Scientific Intelligence, AI and Digital Innovation
Novo Nordisk
{avqg, isdb, rlhb}@novonordisk.com

Abstract

Retrieval Augmented Generation (RAG) has advanced Question Answering (QA) by connecting Large Language Models (LLMs) to external knowledge. However, these systems can still produce answers that are unsupported, lack clear traceability, or misattribute information – a critical issue in the biomedical domain where accuracy, trust and control are essential. We introduce **TripleCheck**, a post-hoc framework that breaks down an LLM's answer into factual *triples* and checks each against both the retrieved context and a biomedical knowledge graph. By highlighting which statements are supported, traceable, or correctly attributed, TripleCheck enables users to spot gaps, unsupported claims, and misattributions, prompting more careful follow up. We present the TripleCheck framework, evaluate it on the SciFact benchmark, analyze its limitations, and share preliminary expert feedback. Results show that TripleCheck provides nuanced insight, potentially supporting greater trust and safer AI adoption in biomedical applications.

1 Introduction

Large Language Models (LLMs) augmented with retrieval, commonly referred to as Retrieval Augmented Generation (RAG), have significantly improved question answering (QA) by grounding responses in external sources. However, despite reducing hallucinations, these systems still exhibit key failures due to inherent system design constraints (Barnett et al., 2024).

In biomedical domains, especially in real-world industry, RAG is relatively underexplored (Bunnell et al., 2025; Ng et al., 2025) but distinct challenges have been pointed out, such as the lack of standard evaluation, unique ethical risks, and recurring problems with irrelevant or misleading information that hamper adoption in a field where both accurate and traceable information is crucial[1]. Additionally, inaccurate or outdated references can compromise the quality of generated responses (Amugongo et al., 2025; Gargari and Habibi, 2025).

Human-AI collaboration research stresses the need for interaction designs that keep users engaged and aware (Song et al., 2025). Without careful explanation mechanisms, users may become overreliant on AI systems (Vasconcelos et al., 2023; Kim et al., 2024; Passi et al., 2024; Zhang et al., 2020). Paradoxically, conventional explanation techniques can increase user trust even when the AI is wrong, elevating the risk of unsubstantiated but plausible-sounding answers (Bansal et al., 2021; González et al., 2021). This underscores the need for new approaches that better surface evidence and improve claim traceability.

To address these gaps and foster appropriate trust in biomedical QA, we propose a post-hoc verification layer that provides fine-grained evidence assessment. Biomedical fact-checking presents unique challenges: knowledge is constantly updated, and contextual nuance often determines the interpretation of evidence (Sosa and Altman, 2022). Overcoming these issues requires strategies that support more nuanced, context-aware evaluations.

We introduce **TripleCheck**, a system-agnostic post-hoc verification framework that can decompose AI-generated biomedical answers into factual triples and checks each for support within both the retrieved context and a large-scale biomedical knowledge graph that aggregates literature, patents, and clinical trials among other sources. This dual approach highlights statements that are supported, traceable, correctly attributed, and flags gaps such as misattributions or conflicting evidence from various sources. This can potentially help users recognize when to be skeptical or seek further evidence to make their own conclusions. By making the support and traceability of claims explicit, TripleCheck

[1] See UN News on WHO's warnings regarding generative AI in healthcare :https://news.un.org/en/story/2023/05/1136707

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 33–47
November 9, 2025 ©2025 Association for Computational Linguistics

aims to calibrate user trust and promote safe reliance on AI answers. Our main contributions are:

- We present **TripleCheck**, a verification framework for biomedical QA that cross-checks answer claims with both retrieved context and a large-scale biomedical knowledge graph.
- We evaluate TripleCheck on a scientific claim verification benchmark (SciFact (Wadden et al., 2020)), showing robust performance against supervised and zero shot alternatives and provides interpretable evidence for each decision. Our analysis shows it disentangles both supported and unsupported information in complex answers.
- We discuss real-world applications and initial expert feedback, illustrating how TripleCheck has the potential to improve trust calibration, transparency, and traceability for workflows such as literature review and clinical QA.

2 Related Work

Scientific Claim Verification Automated fact-checking has progressed from general domains such as political news to specialized areas like biomedicine. Datasets like FEVER (Thorne et al., 2018) supported claim verification against Wikipedia, while SciFact (Wadden et al., 2020) introduced the challenge of verifying scientific claims using abstracts, spurring advances in both evidence retrieval and claim classification. SciFact-Open (Wadden and et al., 2022) broadened this to open-domain settings with over 500,000 abstracts, revealing that scientific evidence is often partial or ambiguous. Other resources have stressed the importance of explainability and evidence alignment for biomedical fact-checking (Kotonya and Toni, 2020; Sarrouti et al., 2021; Saakyan et al., 2021; Kumar et al., 2025).

Beyond traditional claim verification, recent efforts leverage knowledge graphs (KGs) to reduce factual errors, especially given their ability to systematically map relationships among biomedical entities. Notably, recent benchmarks (Lin et al., 2024) challenge AI agents to cross-verify KG-derived facts against the literature, revealing that even advanced LLMs often struggle with this task. Among KG-based approaches, MedGraphRAG (Wu et al., 2024) takes a fundamentally different approach by integrating a knowledge graph directly into the retrieval and generation process, aiming to produce answers that are verified at generation time. In contrast, TripleCheck acts as a post-hoc verification layer: it operates on the output of any generative QA system, requiring no modification or re-training, but instead adding an extra verification step to independently assess claim validity. This distinction means TripleCheck can complement methods like MedGraphRAG by providing an additional safety net.

Other methods propose post-generation claim checking, such as extracting claims from model outputs for KG validation (Guan et al., 2024), or hallucination detection using structured entailment checking over generated answer triples (Sansford et al., 2024). However, these works either do not leverage an external KG for cross-checking (as in (Sansford et al., 2024)), or they lack a user-facing explanation component (as in (Guan et al., 2024)). In contrast, TripleCheck not only combines text entailment and KG validation in a dual-evidence approach, but is also designed with user-understandability and interaction in mind.

Our approach builds on these directions by proposing a zero-shot, post-hoc verification layer that can be added on top of any generative QA system. We uniquely leverage a large-scale biomedical KG to robustly cross-validate atomic answer triples, inspired by recent work in open-domain QA and fact-checking (Li et al., 2025; Kamoi et al., 2023). Importantly, TripleCheck preserves the LLM's original answer, instead surfacing supporting or contradictory evidence for each claim so users can make informed, nuanced judgment – an essential feature in the evolving and often ambiguous landscape of biomedical research.

Trust Calibration and Explainable QA Interfaces Trust calibration – the process by which user trust aligns with the true reliability of an AI system – has emerged as a critical factor in medical AI adoption (Sakamoto et al., 2024). Effective calibration can improve decision accuracy, yet achieving it remains challenging, as trust depends on perceived understandability, technical competence, and system reliability (Darvish et al., 2024). Inadequate calibration, whether overtrust or undertrust, can lead to unsafe outcomes in high-stakes biomedical environments.

There is a growing consensus that AI systems in these domains must support user understanding and oversight through explainable interfaces (Liang and Sonntag, 2025). For example, Li et al. (2024) describe an LLM-assisted QA system with ex-

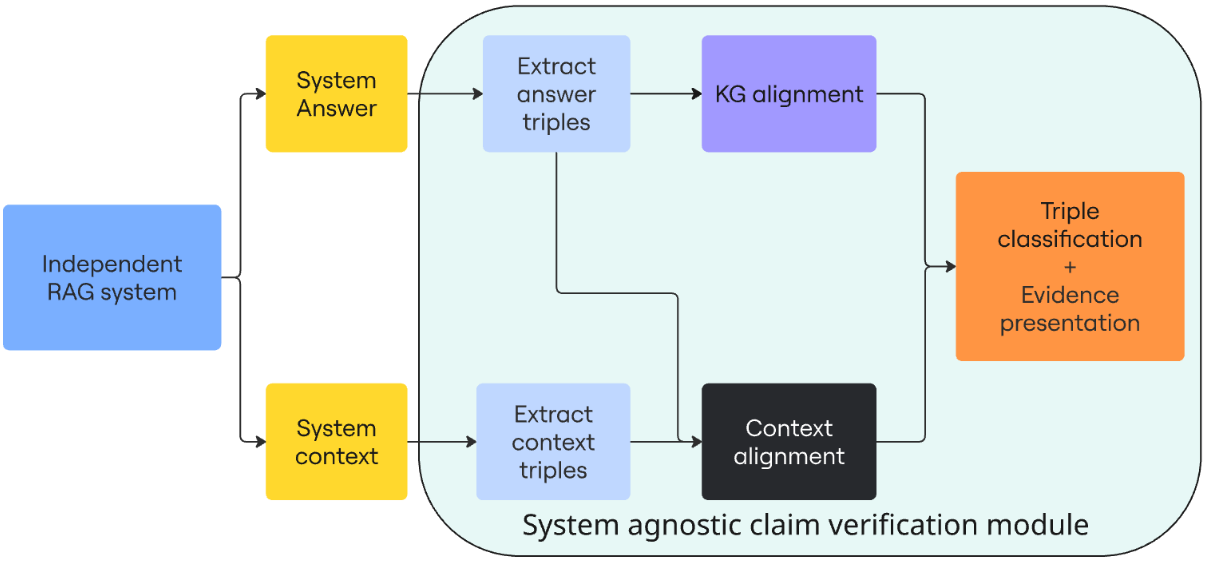


Figure 1: Overview of the TripleCheck pipeline. Given a user question and an answer from a RAG system, TripleCheck extracts atomic triples from the answer (and context) and verifies each one through two channels: (1) alignment with the retrieved context (documents or passages) and (2) cross-checking against a biomedical knowledge graph. Each triple is then labeled as supported, unsupported, or contradicted based on both evidence sources. This claim-level verification can be presented to the user as an interactive interface that highlights which parts of the answer are trustworthy and which require caution.

plicit KG integration for user control, while others caution that some explanations can inadvertently increase overtrust, even when the system is wrong (González et al., 2021; Bansal et al., 2021; Vasconcelos et al., 2023).

Effective interfaces feature interactivity, enabling users to explore not only the answer, but why and how it was produced. This approach helps foster appropriate skepticism and engagement (Rudin et al., 2022; Lai et al., 2023). In the biomedical domain, recent work by Huang et al. (2024) shows that providing multi-hop, interpretable rationales in a drug repurposing model, improved clinicians' accuracy, confidence, and decision efficiency, underscoring the value of transparent, actionable explanations. Similarly, tools such as claim verification with evidence trails (e.g., using SHAP) improve decisions, though risk overreliance without careful design (Liang and Sonntag, 2025).

While we do not fully explore the possibilities of building a sophisticated user interface in this work, TripleCheck is explicitly designed to provide users with the information needed to calibrate trust and promote informed oversight. By breaking down answers into checkable factual units, labeling each as supported, unsupported, or contradicted, and surfacing the underlying evidence from literature or knowledge graphs, TripleCheck offers fine-grained transparency. This enables users to scrutinize each claim with an appropriate level of skepticism or confidence, in line with findings within Human Computer Interaction (HCI) that emphasize user control as fundamental for trust calibration in AI (Passi et al., 2024).

3 Methodology: Post-hoc Claim Verification with TripleCheck

System Overview TripleCheck acts as a post-processor for a standard RAG pipeline. Suppose a user poses a question and the QA system produces an answer along with retrieved documents or passages as context. TripleCheck takes this answer and its supporting context as input, and performs three main steps: (1) Triple Extraction, (2) Evidence Alignment, and (3) Triple Classification. The output is a set of annotated triples derived from the answer, each marked with whether it is supported by the context and/or by the external knowledge graph along with any additional evidence surfaced. Figure 1 illustrates this workflow. As TripleCheck is system-agnostic and never alters the original answers, it can be flexibly added to any QA workflow to provide a second layer of verification.

Triple Extraction The first step breaks each answer into factual triples of the form (Subject, Predicate, Object). For example: from “A deficiency of vitamin B12 increases blood levels of homocysteine, which is a risk factor for heart disease,” we extract *(vitamin B12 deficiency, increases, homocysteine levels)* and *(homocysteine, is a risk factor for, heart disease)*, each treated as an independent claim.

Our method follows recent approaches that combine large language models (LLMs) with post-hoc canonicalization of biomedical entities and relations (Zhang and Soh, 2024). It integrates two main strategies:

- **LLM-based parsing:** We prompt an LLM (GPT-4.1) with instructions (found in the appendix in Table 4, section A.2) to decompose the answer into concise factoid triples (`(Subject, Predicate, Object)`). The prompt is designed to focus on biomedical relations relevant to our KG and domain, and to avoid redundancy or overly broad statements. This captures implicit facts missed by more rigid parsers.
- **NER and RE:** In parallel, a pipeline for Named Entity Recognition (NER) and Relation Extraction (RE) identifies key biomedical entities (e.g., genes, chemicals, diseases) and the relations between them, restricted to a predefined ontology (e.g., “downregulates” , “upregulates”, etc) present in our KG.

Candidate triples from both methods are merged, with further processing to expand abbreviations (e.g., “TNF” → “Tumor Necrosis Factor”) and link entities to KG identifiers. Triples referencing novel or out-of-ontology entities are excluded from KG validation using relations, but retained for textual entailment-based checking. To reduce spurious alignments that could arise during the linking process, an LLM module screens for semantic consistency of the final triples to the system answer. The output is a set of cleaned, distinct factual triples asserted by the answer (see Table 1).

Contextual Evidence Alignment To measure the alignment between the answer and the context, TripleCheck evaluates whether each extracted triple is supported or refuted by the retrieved context. The triple extraction pipeline is also applied to the context documents, yielding sets of context triples for creating a similar structured comparison between claim and context as done by Sansford et al. (2024). For each answer triple, we attempt different matching strategies:

Original Claim	Extracted Triples
Albendazole is used to treat lymphatic filariasis.	(Albendazole, treats, Lymphatic filariasis)
DMRT1 is a sex-determining gene that is epigenetically regulated by the MHM region.	(DMRT1, associated with, sex determination) (MHM region, regulates , DMRT1)
Leukemia associated Rho guanine nucleotide-exchange factor represses RhoA in response to SRC activation.	(Rho guanine nucleotide exchange factor, inhibits, RhoA) (SRC activation, induces, Rho guanine nucleotide exchange factor)

Table 1: Original claims and their extracted triples. Relations and entities are additionally mapped to valid entities and relation types present in our KG.

- **Direct Support:** If the context triples set contains an identical triple to what is in the answer, the claim is marked as explicitly supported by the retrieved context.
- **No Support:** If the triple is absent in the context, it is initially treated as unsupported. However, as absence may result from novel or poorly linked entities, we leverage an LLM to assess if the context entails, contradicts, or is neutral toward the claim (instructions can be found in Table 3, section A.1). Entailment provides implicit support, contradiction triggers a warning, and otherwise the triple remains unsupported.

This strategy allows verification at the individual claim level, revealing when some aspects of an answer are substantiated while others are not.

Knowledge Graph Evidence Alignment TripleCheck simultaneously checks each triple against a biomedical KG that aggregates extracted relationships from sources like PubMed, clinical trials, and patents, among many others. We label extracted triples as:

- **KG-Supported:** If the triple or a suitable variant exists in the KG, we mark it as KG-supported. If the previous is not found, we additionally extract documents mentioning both entities in the triple and run the same textual entailment framework ran during the contextual evidence alignment step to reduce false negatives. We make the supporting evidence available.
- **KG-Contradiction:** If the KG records an opposing assertion (e.g., “A negative cause B” vs. the answer’s “A positive cause B”) via KG

relations or textual entailment method, we flag this as a contradiction and surface the relevant evidence.

- **KG-Unsupported:** If neither support nor contradiction is found, the claim is tagged as unsupported, suggesting either novel science, unsupported assertion or simply a gap in the KG.

Triple Classification By combining contextual evidence and KG-based validation, TripleCheck assigns each claim to one of four main verification categories:

1. **Fully Supported:** Found in both sources, indicating robust scientific consensus and proper attribution.
2. **Supported by KG Only:** Present in the KG but missing from retrieved context, flagging a retrieval or citation gap.
3. **Supported by Context Only:** Found in the context but not in the KG, pointing to possible new concepts or KG incompleteness.
4. **Unsupported:** Unsupported by either evidence channel, raising the possibility of a hallucination or unsubstantiated claim.

Additional flags are included for these cases:

1. **Contradicted in Context:** Explicitly contradicted by at least one retrieved passage, highlighting a likely error in system logic or misleading result.
2. **Contradicted in KG:** Contradicted by the knowledge graph, signaling the existence of contested information.

This fine-grained verification surfaces precisely which portions of an answer are reliable, unsupported, or contested, providing targeted feedback for both users and developers. TripleCheck never alters the original answer; users can decide how to act on verification results, while QA developers may use this information to improve retrieved citations and generation strategies.

Proprietary Components TripleCheck's implementation makes use of certain proprietary components. Specifically, our triple extraction pipeline relies on an in-house biomedical NER and RE system, trained on a broad mix of public biomedical annotations and internal corpora, to achieve wide entity and relation coverage and high accuracy across the biomedical domain. The KG used for evidence alignment is constructed by aggregating structured relationships extracted through automated processes from scientific literature, clinical trial data, patents, and other specialized sources, some of which are not publicly available. An illustrative overview of the knowledge graph structure is shown in Figure 2. While these specific resources cannot be released due to licensing and privacy constraints, the overall TripleCheck framework is system-agnostic and designed for flexibility. Similar pipelines can be constructed using open-source biomedical NER/RE tools and knowledge graphs such as PrimeKG (Chandak et al., 2023). We encourage both academic and industry practitioners to build on or adapt our proposed framework with alternative resources, and view TripleCheck as an inspiration and blueprint for transparent, responsible biomedical QA in both open and proprietary environments.

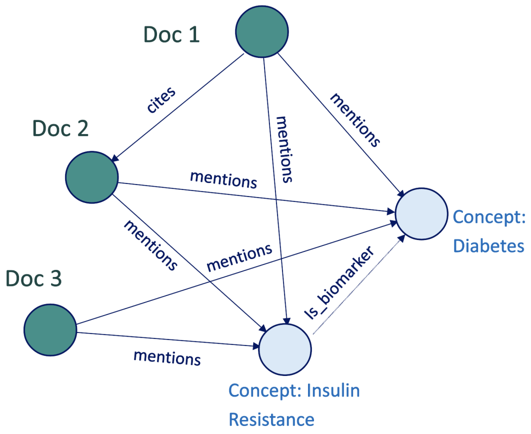


Figure 2: Simplified view of our proprietary KG: unstructured documents contain concept mentions and their relationships. We are able to trace in which documents specific relations are mentioned.

4 Evaluation

We evaluated TripleCheck on the SciFact benchmark (Wadden et al., 2020), where claims are annotated as *Supported*, *Refuted*, or *NEI* (Not Enough Info). As access to the SciFact test set labels is no longer available [2], we perform evaluation on the development set similar to other studies (Deka et al., 2023). While we present results from several other methods on both test set and development set, our

[2] Evaluation on test set was only available via leaderboard which is now closed: `https://leaderboard.allenai.org/scifact/submissions/public`

Model	Precision	Recall	F1
Evaluated on SciFact Dev Set (Zero-Shot Setting)			
TripleCheck (ours)	0.73	0.70	0.70
PubMedBERT-mnli (Deka et al., 2023)	0.66	0.59	0.63
PubMedBERT-mnli-mednli (Deka et al., 2023)	0.84	0.75	0.79
DeBERTa-v3-base-mnli (Deka et al., 2023)	0.42	0.39	0.40
DeBERTa-v3-base-mnli-mednli (Deka et al., 2023)	0.78	0.70	0.74
Evaluated on SciFact Test Set			
Zero-NatVer (Strong et al., 2024) (zero-shot)	-	-	0.55
ClaimGen (entity-based) (Wright et al., 2022)	0.73	0.69	0.71
ClaimGen (BART) (Wright et al., 2022)	0.64	0.79	0.71
MultiVerS (Wadden et al., 2022) (weak-supervision)	0.73	0.71	0.72
VerT5erini (Pradeep et al., 2021)	0.64	0.73	0.68

Table 2: Fact verification results on SciFact. *Top:* All models evaluated on the development set in a zero-shot setting (i.e., **not** fine-tuned on SciFact train data). *Bottom:* Results on the test set, as reported in original publications; including zero-shot and weakly supervised approaches. Note: Due to test set access restrictions, only dev set results are shown for our approach.

key point is that TripleCheck delivers performance broadly in line with state-of-the-art alternatives, highlighting its practical competitiveness.

TripleCheck's output, though more fine-grained, is mapped for comparison: we label the entire data point (a claim from the scifact dataset) as *Supported* if all component triples are at least supported by the retrieved context and none are contradicted, *Refuted* if any triple is contradicted, and *NEI* otherwise. While this mapping is a simplification, it enables comparison on this benchmark.

TripleCheck achieves an F1 of 0.70 on SciFact (dev set) in a zero-shot setting without any task-specific fine-tuning, which is notable given that many comparison models, such as MultiVerS (Wadden et al., 2022) and VerT5erini (Pradeep et al., 2021), are tuned for this task. When comparing against other zero-shot approaches evaluated on the development set (Deka et al., 2023), TripleCheck achieves competitive performance, and outperforms strong baselines. This underscores TripleCheck's out-of-the-box robustness, even though our setup intentionally prioritizes transparency and explainability over strict optimization for SciFact. The results can be seen in Table 2.

Beyond aggregate scores, we also analyzed TripleCheck's outputs for cases where it provides nuanced judgments that classic fact-checkers might miss. We found that about 10% of SciFact's unsupported or contradicted claims were in TripleCheck's *Supported by KG Only* category. Upon inspection, it was evident that while the claim was not supported by the context, the claim was not a non factual claim, and we were able to collect evidence from the biomedical knowledge graph supporting this as an established fact. This reflects a traceability gap, highlighting where a claim may be true even if not cited. Our proposed step for such claims is to improve traceability by fetching additional data, rather than labeling the full claim as non factual. For example, in Scifact, the claim "A deficiency of vitamin B12 increases blood levels of homocysteine." is labeled as unsupported against the context, however, this is a known fact that is well supported in the KG.

4.1 Preliminary Feedback from a Domain Expert

While a full user study was beyond the scope of this research, we solicited preliminary feedback from a biomedical researcher to assess TripleCheck's practical value. The expert reviewed 30 claim-context pairs from the SciFact dataset, along with our system's triple-level evidence (see Figure 3 in the appendix). Feedback was collected via a structured questionnaire and follow up interview. Several key themes emerged:

Granular Verification and Trust Calibration. The domain expert confirmed that decomposing answers into factual triples substantially increased clarity and enabled a more nuanced, calibrated approach to trusting system outputs. Rather than treating each answer as a single unit, the triple-based breakdown highlighted exactly which sub-claims were well-supported, which were missing evidence, and where there was explicit contradiction, echoing prior findings on the value of graph-based and evidence-traceable explanations in medical AI (Johnson et al., 2024). This allowed for more *selective skepticism* according to the expert: reliable portions of an answer could be accepted at face value, while unsupported or contested subsections triggered further review.

Role and Value of Knowledge Graph Support. Feedback emphasized that KG evidence often served as a crucial complement to retrieved context, especially for well-established biomedical facts that may not appear in the narrow selection of retrieved literature. The expert noted that, in practice, when an answer was supported *only* by the KG, they took it as a signal of a gap in retrieval coverage rather than a problem with the claim's validity, pointing to the fact that the user sees the

KG as a more objective and trustworthy source of truth. This aspect of traceability was highly valued for both – confirming canonical domain knowledge and helping efficiently flag true retrieval errors – demonstrating the importance of multi-channel verification over text-only methods. The distinction between KG-backed, context-backed, and unsupported can enable an action-oriented workflow: claims could be triaged for acceptance, additional investigation, or citation gap-filling.

UI Suggestions, and Information Overload. While the expert found the surfacing of supporting evidence to be confidence-boosting, to further reduce cognitive load and speed up review, UI suggestions were made such as: entity highlighting, displaying synonyms, and visually denoting the location of each triple within the evidence. Explanations that grew too detailed or technical could overwhelm non-specialists, consistent with recent findings on explanation overload (Hoffman et al., 2023). The expert also mentioned that as the goals change, the user might be interested in going deep into a topic, while at other times they want to get a high-level overview, therefore, controlling the level of depth and being able to explore and expand based on evidence could be useful. Finally, layered or toggleable presentation and simplified language were highlighted as desirable features.

Gold Standard Inconsistencies and Multiple Verification Channels. The expert occasionally detected that some claims labeled as *Supported* in SciFact were *not substantiated* by the provided abstracts, illustrating limitations of relying on single-source, gold-standard labels. This further supported the premise that multi-evidence verification is necessary to uncover gaps, avoid propagation of citation errors, and empower users to make cautious, context-sensitive decisions.

Taken together, this preliminary expert feedback strongly supports TripleCheck's approach to transparent, claim-level verification across multiple evidence channels. The integration of both KG and literature-derived support increases trust calibration, traceability, and user agency. The decomposition of answers not only aligns with real-world expert workflows but also makes the process of validation more actionable, helping users efficiently accept, investigate, or contest subclaims as needed. Comprehensive, interactive user studies remain a target for future work, but these results demonstrate significant potential for TripleCheck to promote safer and more reliable biomedical AI adoption.

5 Use Cases and Discussion

TripleCheck is broadly applicable to scenarios where users need to trust but verify AI-generated answers. We discuss a few use cases and their potential impact:

Literature Review Assistant: Researchers often use QA systems to quickly summarize findings across papers (e.g., "What causes condition X?"). TripleCheck would allow them to see which claimed causes are well-established (supported by multiple sources or KG) versus which are tentative more contested. It can also reveal if the system's answer includes claims not actually found in any cited papers, prompting the researcher to do a deeper dive for those claims.

Regulatory Document Drafting: In writing reports for drug approval or clinical guidelines, every statement needs a reference. An AI assistant could assist in drafting a section (e.g. drug efficacy) and TripleCheck would immediately flag any statement that lacks backing from the retrieved studies or known medical facts. This helps authors more quickly pinpoint those evidence gaps, saving time and preventing unsubstantiated claims from slipping through.

Clinical Decision Support: A clinician asking an AI assistant about treatment recommendations could benefit from TripleCheck's breakdown. For example, if the answer says "Drug A improves outcome Y and is not associated with side effect Z," TripleCheck might show the first claim is supported by a trial but the second claim is unsupported because the system didn't actually retrieve evidence about side effect Z. The clinician thus knows to be cautious or look up that specific point.

Improving QA System Development: TripleCheck can be used offline by developers of biomedical QA systems to analyze where the system tends to hallucinate or omit citations. If many answers have support only coming from the KG, it may mean the system is relying on prior knowledge not present in the retrieved text — maybe the retrieval component needs improvement. If many answers have "Unsupported" triples, the LLM might be overgeneralizing, suggesting a need for better grounding or post-editing.

Hypothesis Generation: Beyond verification, TripleCheck can assist in hypothesis generation by identifying claims that are plausible yet unsupported by the current evidence base. By inverting the verification output, users can systematically surface statements that are not confirmed in retrieved context or the knowledge graph. These unsupported claims can be then further investigated to see if they highlight potential gaps in scientific knowledge and serve as starting points for novel research questions.

By design, TripleCheck encourages a habit of verification. Rather than replacing human judgment, it guides users to the relevant evidence (or absence thereof). This aligns with the goal of safer deployment of AI in biomedicine: the human expert remains in the loop, making final decisions with a clearer view of the AI's reliability on each sub-point.

6 Limitations and Future Work

While promising, TripleCheck has several limitations:

- **Evaluation is still preliminary**: To date, we lack large-scale studies or professional user testing to validate the usability and benefits of TripleCheck. A crucial and active next step will be conducting a user study to quantitatively evaluate TripleCheck's impact on verification accuracy, confidence, and efficiency, similar to the approach of Huang et al. (2024), who assessed how interpretable explanations improved clinicians' decision-making. The next step is to compare users with and without access to TripleCheck as they assess AI-generated answers, thereby testing whether our framework enhances trust calibration and decision quality. This study will focus on three key outcomes: users' accuracy in claim verification, the time taken for assessment, and their confidence in their decisions.

- **User Experience Considerations**: Highlighting every claim in an answer can lead to information overload and overwhelm users. Careful interface design (e.g., toggleable detail levels) and user training are needed to ensure clarity. Tooltips or onboarding materials could assist users in interpreting verification results. Further exploration on how to build an efficient user interface is an area of future work.

- **Incomplete Knowledge Graph Coverage**: TripleCheck relies on a KG that, while extensive, is not exhaustive. It may lack very recent findings, rare conditions, or new technologies, leading to true claims being labeled as unsupported in KG. Expanding coverage and dynamically updating ontologies could have a positive impact.

- **Triple Extraction Quality**: The accuracy of information extraction directly affects downstream processing. Errors can occur with complex or explanatory sentences, leading to split, merged, or inaccurate triples. While an LLM verification step mitigates some issues, extraction errors can still cause correct claims to be labeled as unsupported and vice versa.

- **Added Latency and Complexity**: The pipeline introduces extra processing (LLM extraction, KG lookup, textual entailment verification) that increases latency. Processing each answer is slower compared to simpler QA systems, and optimizations may be needed for real-time applications.

- **Proprietary Resources**: As previously discussed, various components of TripleCheck are proprietary. While we provide our main TripleCheck system description to support reproducibility, this limitation may hinder exact replication by the research community. As an area of future work, we aim to benchmark public alternatives on fully open resources and encourage efforts to develop analogous public alternatives.

7 Conclusion

We presented **TripleCheck**, a post-hoc verification framework for biomedical QA that decomposes LLM-generated answers into factual triples and verifies each against both retrieved context and a large-scale biomedical knowledge graph. Our SciFact evaluation demonstrated that TripleCheck achieves competitive zero-shot performance while providing fine-grained, interpretable evidence for each claim. Initial expert feedback also suggested that this approach can support more calibrated trust, improve detection of unsupported or contested claims, and aid decision-making in biomedical settings. This initial feedback aligns well with anticipated real-world use, supporting the practical value of TripleCheck in biomedical workflows.

While promising, TripleCheck faces challenges such as refining user interfaces to manage information load, and expanding coverage of supporting knowledge. Most notably, future user studies are necessary to measure TripleCheck's real-world impact on verification accuracy and user confidence.

TripleCheck represents a step toward more transparent, accountable biomedical AI by offering actionable, triple-level evidence to end users and developers. We hope this work encourages further development of evidence-aware QA frameworks, advancing safe and trustworthy use of AI in biomedicine.

Acknowledgments

We are grateful to Zhang Zhong for generously sharing domain expertise and insightful feedback on the practical value of TripleCheck. We also thank colleagues from the Scientific Intelligence team at Novo Nordisk for helpful discussions throughout the development of this work. Their comments and suggestions contributed to refining our approach. Finally, we acknowledge the anonymous reviewers whose thoughtful critiques will shape future iterations of this research.

References

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, CAIN '24, page 194–199, New York, NY, USA. Association for Computing Machinery.

David Bunnell, Mary Bondy, Lucy Fromtling, Emilie Ludeman, and Krishnaj Gourab. 2025. Bridging ai and healthcare: A scoping review of retrieval-augmented generation - ethics, bias, transparency, improvements, and applications.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Mahdieh Darvish, Jan-Hendrik Holst, and Markus Bick. 2024. Explainable ai in healthcare: Factors influencing medical practitioners' trust calibration in collaborative tasks.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2023. Multiple evidence combination for fact-checking of health-related information. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 237–247, Toronto, Canada. Association for Computational Linguistics.

Omid Kohandel Gargari and Gholamreza Habibi. 2025. Enhancing medical ai with retrieval-augmented generation: A mini narrative review. *Digital health*, 11:20552076251337177.

Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do explanations help users detect errors in open-domain qa? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1103–1116.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18126–18134.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257.

Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N Nadkarni, Benjamin S Glicksberg, Nils Gehlenborg, and Marinka Zitnik. 2024. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613.

Ruth Johnson, Michelle M Li, Ayush Noori, Owen Queen, and Marinka Zitnik. 2024. Graph artificial intelligence in medicine. *Annual review of biomedical data science*, 7(2024):345–368.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 822–835.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. 2025. Sciclaimhunt: A large dataset for evidence-based scientific claim verification. *arXiv preprint arXiv:2502.10003*.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1369–1385.

Harry Li, Gabriel Appleby, and Ashley Suh. 2024. Linkq: An llm-assisted visual interface for knowledge graph question-answering. In *2024 IEEE Visualization and Visual Analytics (VIS)*, pages 116–120. IEEE.

Xiangci Li, Sihao Chen, Rajvi Kapadia, Jessica Ouyang, and Fan Zhang. 2025. Minimal evidence group identification for claim verification. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 103–111, Albuquerque, New Mexico. Association for Computational Linguistics.

Siting Liang and Daniel Sonntag. 2025. Explainable biomedical claim verification with large language models. *arXiv preprint arXiv:2502.21014*.

Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z Li, and Kaicheng Yu. 2024. Biokgbench: A knowledge graph checking benchmark of ai agent for biomedical science. *arXiv preprint arXiv:2407.00466*.

Karen Ka Yan Ng, Izuki Matsuba, and Peter Chengming Zhang. 2025. Rag in health care: A novel framework for improving communication and decision-making by addressing llm limitations. *NEJM AI*, 2(1):AIra2400380.

Samir Passi, Shipi Dhanorkar, and Mihaela Vorvoreanu. 2024. Appropriate reliance on generative ai: Research synthesis. Technical report, Technical Report MSR-TR-2024-7. Microsoft. https://www. microsoft. com/en-us ....

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific claim verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103, online. Association for Computational Linguistics.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Tetsu Sakamoto, Yukinori Harada, Taro Shimizu, and 1 others. 2024. Facilitating trust calibration in artificial intelligence–driven diagnostic decision support systems for determining physicians' diagnostic accuracy: Quasi-experimental study. *JMIR Formative Research*, 8(1):e58666.

Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. 2024. Grapheval: A knowledge-graph based llm hallucination evaluation framework. *arXiv preprint arXiv:2407.10793*.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jaeyoon Song, Zahra Ashktorab, Qian Pan, Casey Dugan, Werner Geyer, and Thomas W Malone. 2025. Interaction configurations and prompt guidance in conversational ai for question answering in human-ai teams. *arXiv preprint arXiv:2505.01648*.

Daniel N Sosa and Russ B Altman. 2022. Contexts and contradictions: a roadmap for computational drug repurposing with knowledge inference. *Briefings in bioinformatics*, 23(4):bbac268.

Marek Strong, Rami Aly, and Andreas Vlachos. 2024. Zero-shot fact verification via natural logic and large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17021–17035, Miami, Florida, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38.

David Wadden and et al. 2022. Scifact-open: Towards open-domain scientific claim verification. In *Findings of EMNLP 2022*, pages 347–359.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2596–2611.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An LLM-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836, Miami, Florida, USA. Association for Computational Linguistics.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305.

A Example prompts

A.1 Entailment/contradiction prompt

In Table 3, we show the prompt used for assessing textual entailment at different stages. We used the same prompt to verify final triples are aligned with system answer, to verify the triple is aligned with the context and to verify that the triple is aligned to any external evidence we have found via the KG.

A.2 Triple extraction

In Table 4, we show the instructions used for extracting initial triples. The initial triples were additionally linked to ontology terms using our proprietary entity linking system, and were afterwards verified against the actual claim to ensure consistency in the final triples set.

B Expert feedback questionnaire

Figure 3 shows how evidence was initially presented to the expert for initial feedback on what could make the result more useful. In Table 5, we have additionally compiled some of the key comments coming both from the written feedback and interview categorized into themes.

System Prompt for LLM-based textual entailment

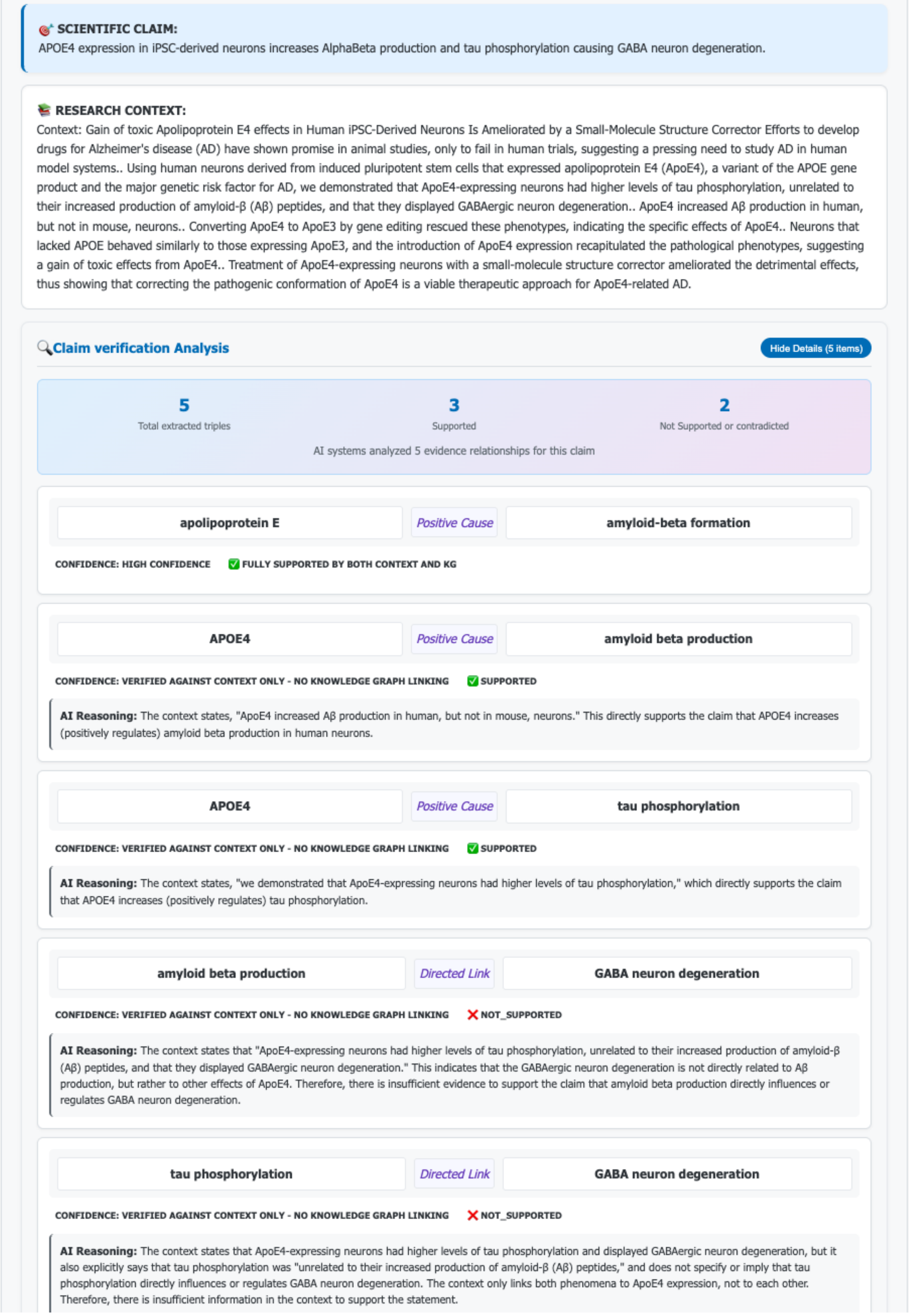
```
You are a claim-verification system.
      Your task is to determine
     whether the given statement is
     supported (directly, indirectly
     , or can be reasonably inferred
     , even if this requires
     combining context and general
     biological knowledge) by the
     provided context.

CONTEXT:
{context}

STATEMENT TO VERIFY:
"{statement}"

VERIFICATION RULES:
1. Answer "YES" if the statement is
     supported by the context, can
     be logically inferred from the
     context, **or if it is
     biologically plausible and
     consistent with accepted
     scientific background knowledge
     .** You can accept reasonable
     combinations of entities as
     long as the overall logic is
     supported, even if not every
     link is explicitly present in
     the context. Do not be overly
     strict about requiring explicit
      verbatim phrasing or full
     mechanistic details **favor a
      positive answer if the overall
      claim is well-supported or
     reasonably implied.**
2. Answer "CONTRADICTION" if the
     statement clearly contradicts
     the context.
3. Answer "NO" only if there is
     insufficient information, the
     claim is irrelevant, or
     biological plausibility is
     seriously lacking or unclear.

RESPONSE FORMAT:
Begin with "YES", "CONTRADICTION",
     or "NO" on its own line. Do not
      start in any other way.
Then provide a brief, evidence-based
      explanation that quotes or
     paraphrases relevant portions
     of the context and/or uses well
     -accepted biological background
      if relevant.

YOUR VERIFICATION:
```

Table 3: System prompt for textual entailment as used in this work.

System Prompt for LLM-based Triple Extraction

```
You are an expert extracting entities and relations from scientific text.

    Given an answer to a scientific question, extract the claims in triples format.

    Your output must be a valid JSON array containing exactly one object per triple in this format:
    [["subject1", "relation1", "object1"], ["subject2", "relation2", "object2"], ...]
    **CRUCIAL RULES     READ CAREFULLY:**
    1. Do NOT use intervention phrases, experimental treatments, or contextual language as entities:
        - Disallow: "PARN targeting", "PARN inhibition", "knockout of PARN", "overexpression of X", "
             activation of Y"
        - Allow only: the core biological entity/process itself (e.g., "PARN", "TP53", "insulin
             maturation")

    2. Use concise, ontology-friendly names (2-4 words max), established biomedical terms, no
         abbreviations unless standard.

    3. DO NOT encode intervention or experiment type in subject/object. NEVER use experimental
         manipulation phrases as entities.

    - Do not use long descriptive phrases or qualifiers as entities.
    - Use 2-4 words maximum for each entity, and keep them concise and ontology-friendly.
    - Use established gene names, protein names, disease terms, and biological processes if possible.

    **Direct Entity-Relation-Entity Guidance:**
    - PREFERRED: ['Gene X', 'Directed Link', 'Process Y']
    - AVOID:    ['Knockdown of Gene X', 'Directed Link', 'Upregulated Process Y']
    - Do not build effects (like "loss", "increase", or "compromised state") into the entity. Use the
          proper relation instead.

    **Decompose Complex Entities:**
    - Break up complex cause-effect phrases into multiple, simpler, functionally meaningful triples
         only using entities present in standard biomedical ontologies.

    **Relation Types (use only these) but keep in mind the mentioned above:**
        -Focus mostly on these
        Directed Link: Direct interaction between entities. Can include correlations or associations.
              IMPORTANT When in doubt of direction, use this.
        Negative Cause: Causes a decrease or inhibition in the target entity.
        Not Directed Link: Interaction without specified direction.
        Not Negative Cause: Does not lead to a negative effect.
        Not Positive Cause: Does not lead to a positive effect.
        Positive Cause: Causes an increase or stimulation in the target entity.

        PPI (Protein-Protein Interaction): Interaction affecting protein function.
        DDI (Drug-Drug Interaction): Interaction affecting drug effectiveness.

        - These can be used to but limit them
        ACTIVATOR: Increases activity of a process or molecule.
        AGONIST: Initiates response by combining with a receptor.
        AGONIST-ACTIVATOR: Initiates and enhances activity.
        AGONIST-INHIBITOR: Acts as agonist and inhibitor.
        ANTAGONIST: Inhibits physiological action of another.
        DIRECT-REGULATOR: Directly modulates target activity.
        INDIRECT-DOWNREGULATOR: Indirectly decreases target activity.
        INDIRECT-UPREGULATOR: Indirectly increases target activity.
        INHIBITOR: Slows or prevents chemical reactions.
        PART-OF: Entity is a component of a larger structure.
        PRODUCT-OF: Entity is a result of a process.
        SUBSTRATE: Molecule acted upon by an enzyme.
        SUBSTRATE_PRODUCT-OF: Substrate converted into a product.
        undefined: Relationships not yet characterized or classified in this list but are still valid
             .

    - For abbreviations, prefer the full name if confidently available from context.
    - Both subject and object must be concise entities/concepts, not specific statements, modifiers,
         or experimental constructs.
    - Do NOT repeat triples (even if synonyms are used in the text).
    - If none of the relations are present, use "undefined". Do NOT invent new relations.
    - If there are no triples present, return [].
    - Your output must be valid JSON directly parsable by `json.loads()` as a list of triple lists (
         not nested or with extra structure)  e.g. [["subject1", "Directed Link", "object1"], ["
         subject2", "Part-Of", "object2"]].
    - Do NOT include any explanations or text outside the JSON array.
```

Table 4: System prompt for triples extraction as used in this work.



🎯 **SCIENTIFIC CLAIM:**
APOE4 expression in iPSC-derived neurons increases AlphaBeta production and tau phosphorylation causing GABA neuron degeneration.

📚 **RESEARCH CONTEXT:**
Context: Gain of toxic Apolipoprotein E4 effects in Human iPSC-Derived Neurons Is Ameliorated by a Small-Molecule Structure Corrector Efforts to develop drugs for Alzheimer's disease (AD) have shown promise in animal studies, only to fail in human trials, suggesting a pressing need to study AD in human model systems.. Using human neurons derived from induced pluripotent stem cells that expressed apolipoprotein E4 (ApoE4), a variant of the APOE gene product and the major genetic risk factor for AD, we demonstrated that ApoE4-expressing neurons had higher levels of tau phosphorylation, unrelated to their increased production of amyloid-β (Aβ) peptides, and that they displayed GABAergic neuron degeneration.. ApoE4 increased Aβ production in human, but not in mouse, neurons.. Converting ApoE4 to ApoE3 by gene editing rescued these phenotypes, indicating the specific effects of ApoE4.. Neurons that lacked APOE behaved similarly to those expressing ApoE3, and the introduction of ApoE4 expression recapitulated the pathological phenotypes, suggesting a gain of toxic effects from ApoE4.. Treatment of ApoE4-expressing neurons with a small-molecule structure corrector ameliorated the detrimental effects, thus showing that correcting the pathogenic conformation of ApoE4 is a viable therapeutic approach for ApoE4-related AD.

🔍 **Claim verification Analysis**

5	3	2
Total extracted triples	Supported	Not Supported or contradicted

AI systems analyzed 5 evidence relationships for this claim

apolipoprotein E	*Positive Cause*	**amyloid-beta formation**

CONFIDENCE: HIGH CONFIDENCE ✅ FULLY SUPPORTED BY BOTH CONTEXT AND KG

APOE4	*Positive Cause*	**amyloid beta production**

CONFIDENCE: VERIFIED AGAINST CONTEXT ONLY - NO KNOWLEDGE GRAPH LINKING ✅ SUPPORTED

AI Reasoning: The context states, "ApoE4 increased Aβ production in human, but not in mouse, neurons." This directly supports the claim that APOE4 increases (positively regulates) amyloid beta production in human neurons.

APOE4	*Positive Cause*	**tau phosphorylation**

CONFIDENCE: VERIFIED AGAINST CONTEXT ONLY - NO KNOWLEDGE GRAPH LINKING ✅ SUPPORTED

AI Reasoning: The context states, "we demonstrated that ApoE4-expressing neurons had higher levels of tau phosphorylation," which directly supports the claim that APOE4 increases (positively regulates) tau phosphorylation.

amyloid beta production	*Directed Link*	**GABA neuron degeneration**

CONFIDENCE: VERIFIED AGAINST CONTEXT ONLY - NO KNOWLEDGE GRAPH LINKING ❌ NOT_SUPPORTED

AI Reasoning: The context states that "ApoE4-expressing neurons had higher levels of tau phosphorylation, unrelated to their increased production of amyloid-β (Aβ) peptides, and that they displayed GABAergic neuron degeneration." This indicates that the GABAergic neuron degeneration is not directly related to Aβ production, but rather to other effects of ApoE4. Therefore, there is insufficient evidence to support the claim that amyloid beta production directly influences or regulates GABA neuron degeneration.

tau phosphorylation	*Directed Link*	**GABA neuron degeneration**

CONFIDENCE: VERIFIED AGAINST CONTEXT ONLY - NO KNOWLEDGE GRAPH LINKING ❌ NOT_SUPPORTED

AI Reasoning: The context states that ApoE4-expressing neurons had higher levels of tau phosphorylation and displayed GABAergic neuron degeneration, but it also explicitly says that tau phosphorylation was "unrelated to their increased production of amyloid-β (Aβ) peptides," and does not specify or imply that tau phosphorylation directly influences or regulates GABA neuron degeneration. The context only links both phenomena to ApoE4 expression, not to each other. Therefore, there is insufficient information in the context to support the statement.

Figure 3: User feedback interface. While the intended use is in an interactive QA setting, this preliminary study presented the interface in a static, questionnaire format to collect initial expert feedback. Future work will focus on exploring different presentation formats and interactive modes.

Theme: Evidence & Explanation Quality
“The explanation directly referenced the supporting evidence, which was helpful.”
“The additional reasoning summarized relevant points well, presented additional evidence and matched my interpretation.”
“Sometimes the explanation focused too much on specific concepts, making it less broadly useful.”
“Having both ‘supported’ and ‘contradicted’ reasoning was logical; it’s important to consider context, while ultimately I would say this is a supported claim, as an expert I have similar concerns as the contradictions surfaced.”
“If I feel an answer is incomplete or uncertain, I’ll ask for more detail or reasoning before accepting it.”
“If the context doesn’t really support the claim, I become wary and might not trust that part of the answer. So having the additional evidence is key”
Theme: User Interface & Usability
“Highlighting the triple location or the keywords like gene names or important biological terms would help me quickly locate evidence in the text.”
“Some explanations were overly technical or as difficult to follow as the original literature. It could be nice to choose how deep to go yourself”
“It would help to see synonyms of entities or have key parts of the triple highlighted directly in the evidence.”
“Claims can be hard to understand if you are not an expert in the topic, simplified breakdowns or highlights would make it easier.”
Theme: Knowledge Graph (KG) Value
“KG support was very useful, especially when the retrieved context didn’t cover established facts.”
“Recognizing when information is canonical, even if not in the provided context, adds confidence.”
“I often trust facts from the KG more, especially when the answer is missing context evidence, it gives reassurance about general scientific truth.”
“Sometimes, the KG picked up on a missing fact from the literature, and that signaled an issue with the context rather than a problem with the claim itself.”
Theme: Exploration & Workflow
“If most triples are supported, I move on, if any aren’t, I dig deeper or ask for more sources. Seeing the breakdown helps me focus.”
“Having access to more detailed evidence when I want it, without being overwhelmed would make deciding whether claims are true or not easier.”
Theme: Areas for Improvement
“Going through the retrieved context can already be complex, so have a simplified language in the breakdown would be helpful.”
“Balance between pointing out specifics and giving a general overview in reasoning. User should be able to choose how deep to go into the details”

Table 5: Sample categorized expert feedback (paraphrased) from the TripleCheck evaluation.