allsci
https://allsci.com

Reliable Hypothesis Extraction with Generate-Judge-Verify

AUTHORS

Akhil Pandey Akella
Allsci Corp, Sunwater Capital

ACKNOWLEDGEMENT

This work is proprietary and all artifacts are owned by Allsci Corp.

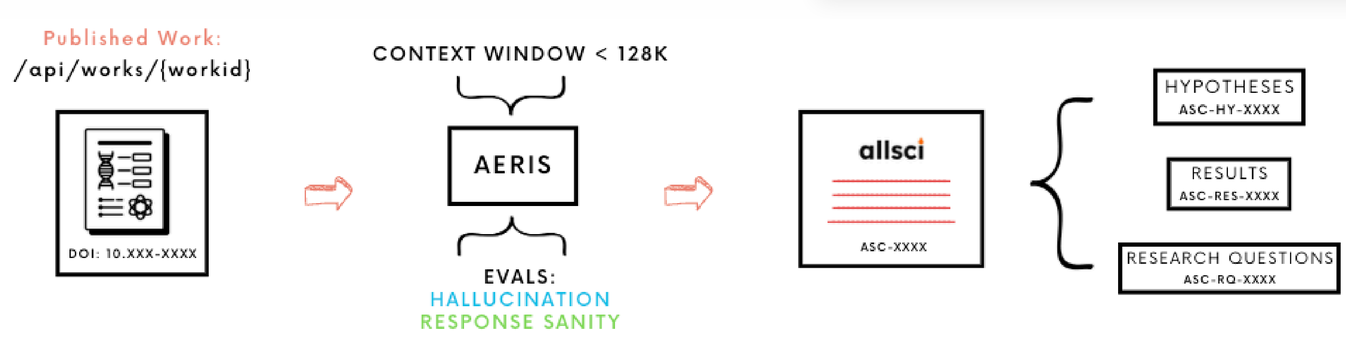
*Hypothesis extraction is attractive for science-of-science because it promises machine-readable maps of what papers propose into **ATOMS** (**hypotheses, results, research questions**) before those claims are tested at scale.*

In so far as a scientific statement speaks about reality, it must be falsifiable: and in so far as it is not falsifiable, it does not speak about reality."
— Karl R. Popper, The Logic of Scientific Discovery

Broader Objective

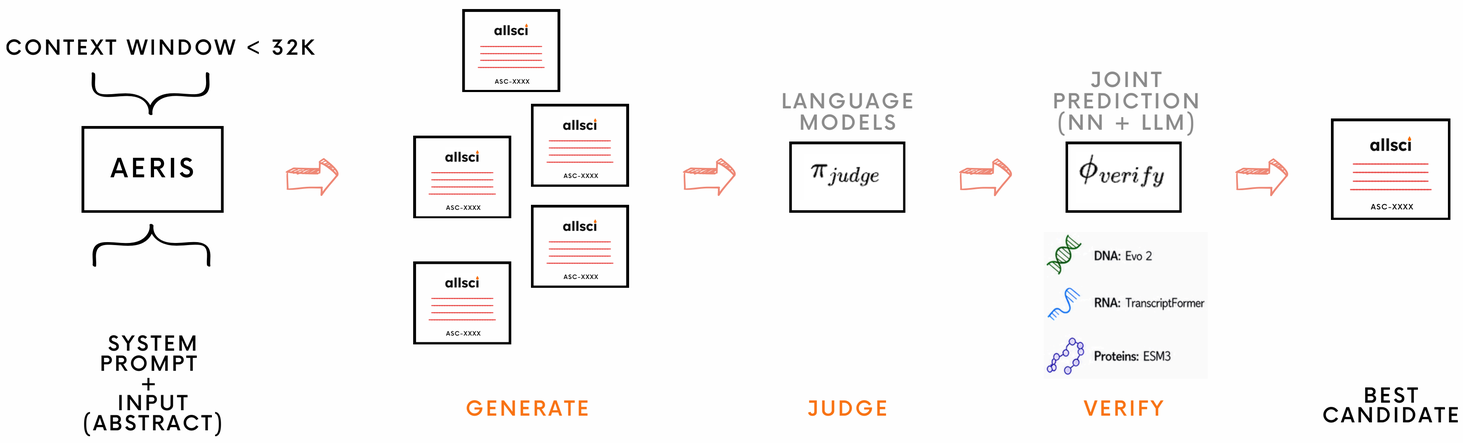
Scientific works are a dense source of information with several signals on the principle contributions of the work. Hypothesis extraction is attractive for science-of-science because it promises machine-readable maps of what papers propose before those claims are tested at scale. In practice, however, the bottleneck is not merely finding plausible hypotheses; it is producing outputs that are structurally valid, evidentially grounded, and disciplined enough to abstain when the abstract does not actually support extraction. Early versions of \texttt{hypogen} showed that simply providing more context did not solve this problem: both abstract-only and focused full-text settings often returned malformed structures or specific-sounding claims untethered from the source text. We therefore reframed the task as \emph{reliability-constrained, abstract-grounded claim extraction} rather than open-ended summarization.

Fig 1: Faithful hypotheses extraction from academic articles



$$\mathrm{Hall}(y,\tilde{a}) = 1 - \frac{1}{|Q(y)|} \sum_{q \in Q(y)} \mathbf{1}[q \subseteq_n \tilde{a}],$$

Fig 2.a: Generate, Judge and Verify framework for resource constrained extraction



AERIS takes a given scientific abstract, generates multiple candidate hypotheses within a compact context window and uses the candidate hypotheses generated by language models, then ranked by an LLM-based judge.

The strongest candidates are passed to a verification layer that combines neural predictors with language-model reasoning, including biological foundation models for DNA, RNA, and protein-level evidence.

This joint generate–judge–verify workflow as seen in Fig 2.a) filters broad hypothesis space into the most biologically plausible candidate.

The AllSci Evidence Engine connects generated hypotheses to supporting or contradicting evidence from scientific literature.

Given an AllSci work and an associated hypothesis, the engine retrieves relevant article-level evidence and classifies the relationship as support, refute, or mixed.

As shown in Fig 2.b) the output links hypotheses, results, and source publications into an evidence graph, making each hypothesis traceable to the studies that test or contextualize it.

Fig 2.b: Evidence engine outlining supporting, refuting and mixed evidence for generated hypotheses

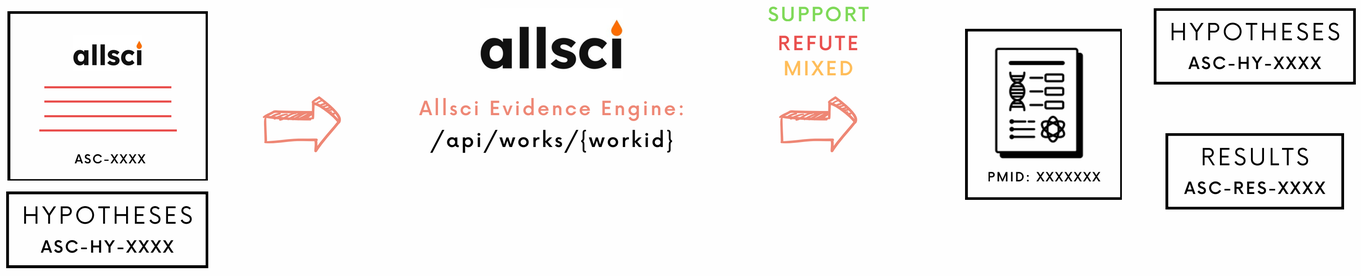


Fig3: Representations of Hypotheses