Universität Zürich^UZH

Dynamic and Distributed Information Systems

Ai2

EMNLP 2025
Suzhou, China | 中国苏州

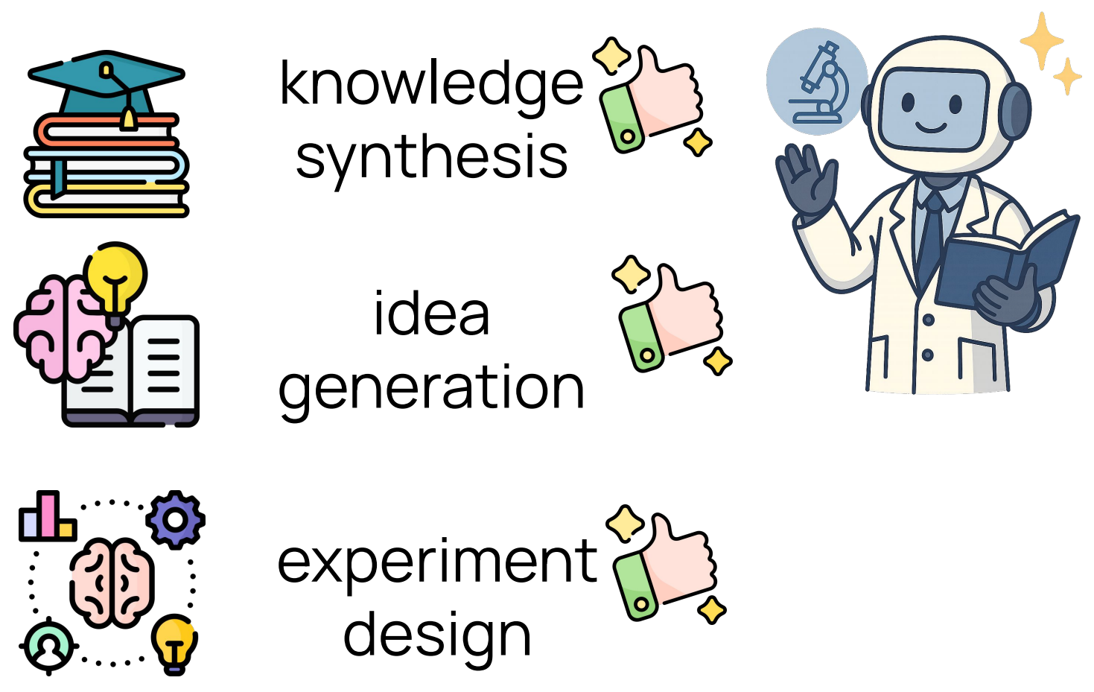
HypER : Literature-grounded Hypothesis Generation and Distillation with Provenance

Rosni Vasu, Chandrayee Basu, Bhavana Dalvi Mishra,
Cristina Sarasua, Peter Clark, Abraham Bernstein
University of Zurich, Cornell University, Allen Institute for AI

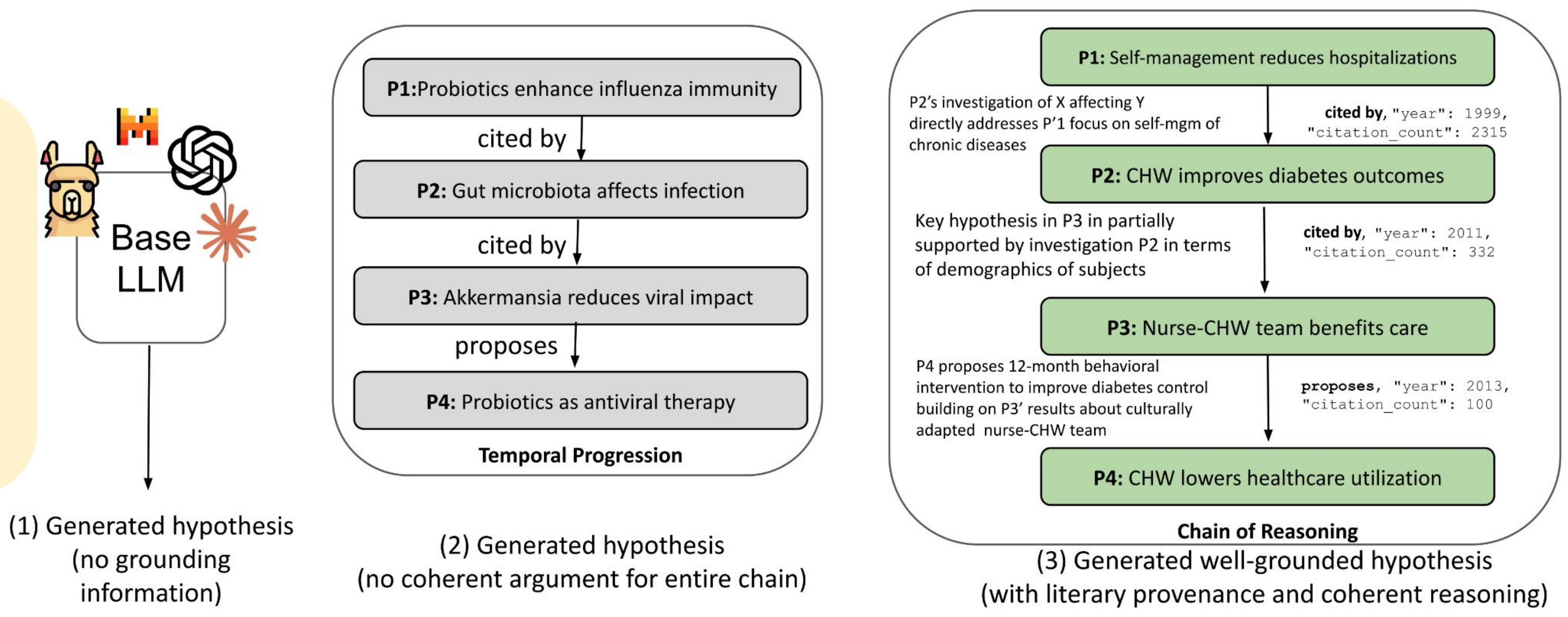
HypER Code

HypER PDF

Motivation



- LLMs can act as AI scientists, generating hypotheses and designing experiments.
- But scientific hypotheses must be logically grounded, interpretable, and based on existing knowledge.


Base LLM
P1:Probiotics enhance influenza immunity
cited by
P2: Gut microbiota affects infection
cited by
P3: Akkermansia reduces viral impact
proposes
P4: Probiotics as antiviral therapy
Temporal Progression
P1: Self-management reduces hospitalizations
P2's investigation of X affecting Y directly addresses P'1 focus on self-mgm of chronic diseases
cited by, "year": 1999, "citation_count": 2315
P2: CHW improves diabetes outcomes
Key hypothesis in P3 in partially supported by investigation P2 in terms of demographics of subjects
cited by, "year": 2011, "citation_count": 332
P3: Nurse-CHW team benefits care
P4 proposes 12-month behavioral intervention to improve diabetes control building on P3' results about culturally adapted nurse-CHW team
proposes, "year": 2013, "citation_count": 100
P4: CHW lowers healthcare utilization
Chain of Reasoning

(1) Generated hypothesis (no grounding information)

(2) Generated hypothesis (no coherent argument for entire chain)

(3) Generated well-grounded hypothesis (with literary provenance and coherent reasoning)

Existing work:

- Reliance on co-occurrence patterns or surface-level similarity
- Poor logical progression and no provenance between ideas

⚠ In biomedical research, hypotheses must be grounded in evidence, logically interpretable, and scientifically valid → not just creative.

How can we train an LLM to navigate the noisy literature and generate novel and impactful ideas that are grounded in a solid understanding of existing work?

HypER: Our Approach

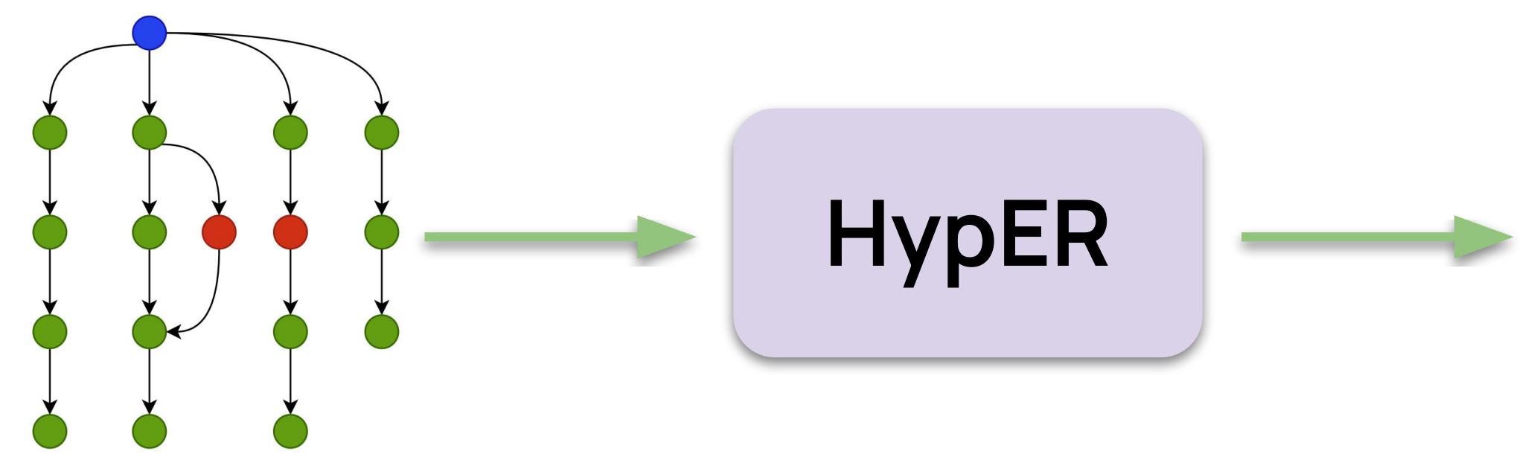
HypER is fine-tuned jointly on three complementary tasks to capture reasoning from local dependencies to multi-hop chains

One-hop Relevance Classification (1-hop)
Input: source paper + target paper
Output: fine-grained relevance score → {0: irrelevant, 1: inspired, 2: dependent}

Multi-hop Agnostic Chain Validation (multi-hop-A)
Input: temporally ordered paper chain
Output: valid / invalid + breakpoints if invalid

Multi-hop Contextual Chain Validation (multi-hop-C)
Input: paper chain + target hypothesis
Output: valid / invalid + breakpoints

Hypothesis Generation with HypER



```
Analysis:...
Rationale:...
Research Idea:...
Hypothesis: In patients with critical limb
ischemia due to infrapopliteal artery
disease, drug-eluting BVS will result in
higher primary patency rates, lower rates
of major adverse limb events, and improved
limb salvage rates compared to angioplasty
at 1 year.,.....
```

Research Questions

RQ1. Can HypER differentiate between valid and invalid reasoning chains?

RQ2. Does reasoning chain validation improve the quality of generated hypotheses?

Dataset and Models

- **Models:** Phi-3-3.8B, LLaMA-3.2-3B, MistralLite-7B-32K
- Chosen for similar size, instruction-tuning, and long-context handling
- Baselines: no reasoning-chain supervision
- HypER refers to the fine-tuned version of the Phi-3-mini-128k-instruct-3.8B model
- Chosen as main model – performed best among all tested SLMs

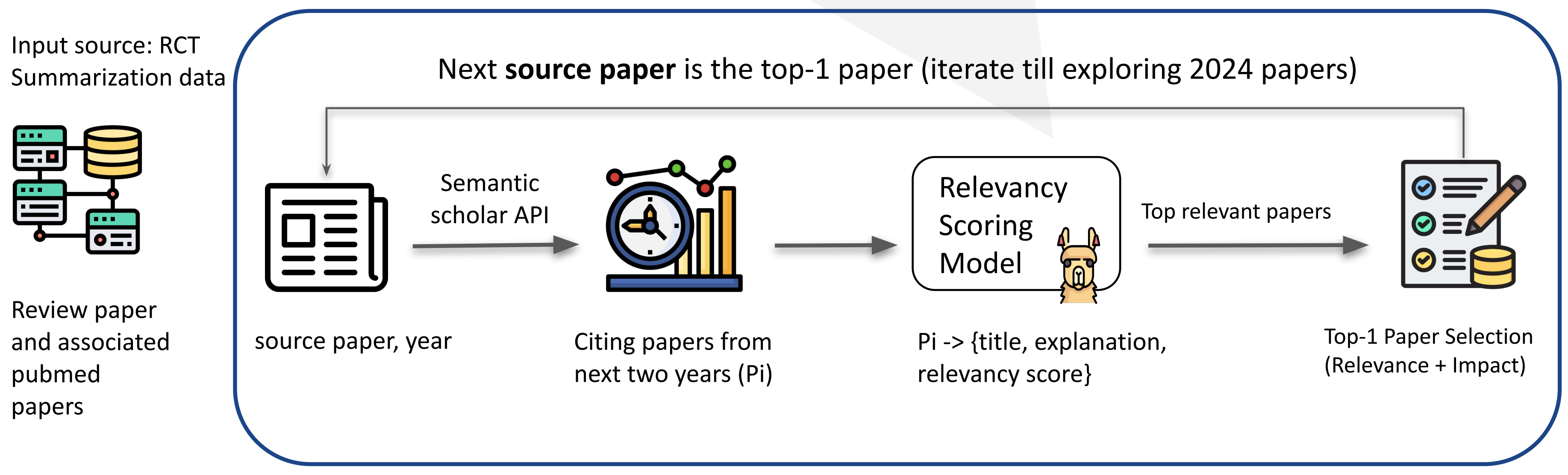
Building Temporal Reasoning Chains

Curated temporal reasoning chains form the backbone of HypER

```
0: No connection to source or a review paper

1: Inspired by the hypothesis or findings of the source paper

2: Partially dependent on the findings of the source paper (contains
sub-hypotheses)
```

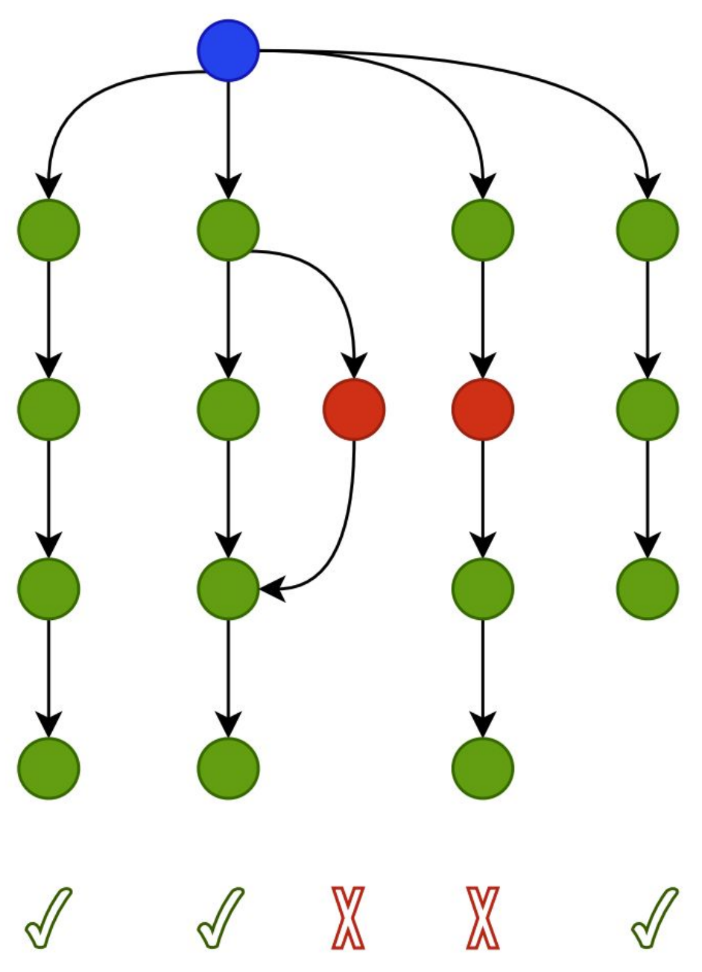

Input source: RCT Summarization data
Review paper and associated pubmed papers
Next source paper is the top-1 paper (iterate till exploring 2024 papers)
Semantic scholar API
Relevancy Scoring Model
Top relevant papers
source paper, year
Citing papers from next two years (Pi)
Pi -> {title, explanation, relevancy score}
Top-1 Paper Selection (Relevance + Impact)

✓ **Valid chains:**

➔ directly obtained from the iterative construction process (379 chains)

✗ **Negative chains:**

➔ Easy: swap with irrelevant paper (1184)
➔ Hard: break or insert incoherent links (455)



TakeAways

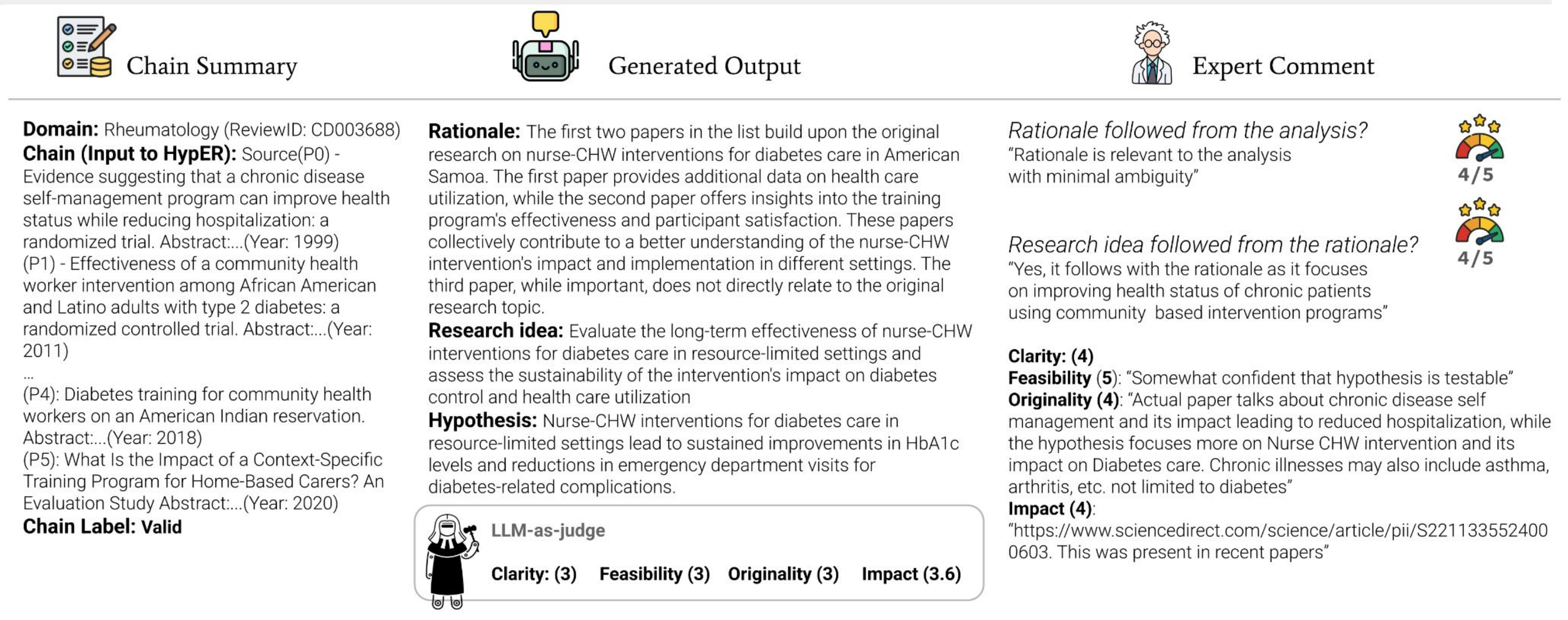
- **+22%** absolute F1 over base models in distinguishing valid vs. invalid reasoning chains.
- Generates more evidence-grounded hypotheses **(0.327 vs. 0.305 baseline)**.
- Human experts (clinicians & biomedical researchers) rated **outputs > 3.5 / 5 for feasibility and impact**, in some cases anticipatory of recent studies.

Example of a hypothesis generated by HypER with expert review

Semantic similarity ≠ reasoning quality

PubMedBERT scores are nearly identical for valid and invalid chains (~0.988 vs. 0.987), proving that semantic similarity alone cannot ensure logical coherence.

- HypER aligns with valid parts of partially invalid chains, showing deeper reasoning awareness.
- Captures scientific dependencies and differentiates valid vs. invalid reasoning chains beyond surface similarity.
- Generates grounded hypotheses and enables evidence-driven, structured research ideation.



Chain Summary

Domain: Rheumatology (ReviewID: CD003688)
Chain (Input to HypER): Source(P0) - Evidence suggesting that a chronic disease self-management program can improve health status while reducing hospitalization: a randomized trial. Abstract:...(Year: 1999)
(P1) - Effectiveness of a community health worker intervention among African American and Latino adults with type 2 diabetes: a randomized controlled trial. Abstract:...(Year: 2011)
...
(P4): Diabetes training for community health workers on an American Indian reservation. Abstract:...(Year: 2018)
(P5): What Is the Impact of a Context-Specific Training Program for Home-Based Carers? An Evaluation Study Abstract:...(Year: 2020)
Chain Label: Valid

Generated Output

Rationale: The first two papers in the list build upon the original research on nurse-CHW interventions for diabetes care in American Samoa. The first paper provides additional data on health care utilization, while the second paper offers insights into the training program's effectiveness and participant satisfaction. These papers collectively contribute to a better understanding of the nurse-CHW intervention's impact and implementation in different settings. The third paper, while important, does not directly relate to the original research topic.
Research idea: Evaluate the long-term effectiveness of nurse-CHW interventions for diabetes care in resource-limited settings and assess the sustainability of the intervention's impact on diabetes control and health care utilization
Hypothesis: Nurse-CHW interventions for diabetes care in resource-limited settings lead to sustained improvements in HbA1c levels and reductions in emergency department visits for diabetes-related complications.

LLM-as-judge
Clarity: (3) **Feasibility (3)** **Originality (3)** **Impact (3.6)**

Expert Comment

Rationale followed from the analysis?
"Rationale is relevant to the analysis with minimal ambiguity"
4/5

Research idea followed from the rationale?
"Yes, it follows with the rationale as it focuses on improving health status of chronic patients using community based intervention programs"
4/5

Clarity: (4)
Feasibility (**5**): "Somewhat confident that hypothesis is testable"
Originality (4): "Actual paper talks about chronic disease self management and its impact leading to reduced hospitalization, while the hypothesis focuses more on Nurse CHW intervention and its impact on Diabetes care. Chronic illnesses may also include asthma, arthritis, etc. not limited to diabetes"
Impact (4):
"https://www.sciencedirect.com/science/article/pii/S2211335524000603. This was present in recent papers"